ABSTRACT

CHHABRA, BANPREET SINGH. Quality Control of Crowd Labeling for Improving Training Data for Machine Learning. (Under the direction of Dr. Edward F. Gehringer).

The effectiveness of machine learning depends on the quality of the training data. In some fields, the training data is generated from crowdsourced labels. Ensuring the reliability and accuracy of crowd-generated labels is a critical challenge. To address this, we present comprehensive research to develop and implement robust quality control strategies in crowd labeling. We focus on enhancing the effectiveness of feedback and suggestions by evaluating the taggers and the quality of tags they assign by using natural language processing techniques and machine learning. Our approach aims to assign reliability metrics to each tagger and tag, enabling researchers to filter and create machine-learning training datasets from the most reliable annotations.

This research addresses this issue by implementing four quality control strategies in the domain of peer assessment and comparing their performance with manual grade scores assigned by professors. The four key quality control strategies encompass identifying taggers who tag too quickly, detecting taggers providing inconsistent labels, uncovering unreliable taggers employing pattern-based tagging, and performing agreement/disagreement analysis for tags. By individually implementing and assessing the impact of each strategy on data alignment with manual grade scores, we ascertain their effectiveness in enhancing the reliability of crowd-generated labels.

While our research is tailored to peer assessment, its implications span various domains relying on crowd labeling for data annotation. The empirical evidence provided in this study guides the selection and implementation of suitable quality-control strategies, ultimately elevating the reliability and accuracy of crowd labeling in diverse applications. By adopting the insights from this research, researchers leveraging crowd-sourced data can optimize their efficiency, accuracy, and decision-making processes. This study contributes to advancing crowd labeling practices and encourages further exploration of automated assessment methodologies.

Quality Control of Crowd Labeling for Improving Training Data for Machine Learning

by
Banpreet Singh Chhabra

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Computer Science

Raleigh, North Carolina
2023

APPROVED BY:

Dr. Collin Lynch	Dr. Noboru Matsuda

Dr. Edward F. Gehringer
Chair of Advisory Committee

ACKNOWLEDGEMENTS

I am extremely grateful for the invaluable advice and perpetual encouragement extended by my supervisor/advisor Dr. Edward F. Gehringer from the preliminary stages till the final completion of the project. I am also thankful to my colleagues and other TAs from computer science and other departments for providing me with unbiased feedback to surveys during my research. I would like to thank the department of Computer Science at North Carolina State University for assisting me in my research and non-research activities.

I would like to extend my thanks to the following people without whom it would not have been possible for me to achieve this work. I owe my sincere thanks to my advisor, for his all time support and guidance. I am very grateful to him for his valuable time, continuous advice, and his feedback throughout this thesis that helped me to improve my work. I learned a lot from our weekly meetings and his insights helped me to generate creative ideas in my mind. I would like to thank my other committee members, Dr. Collin Lynch, for helping me to work in the right direction and for all his guidance and discussions throughout my thesis work. Dr. Noboru Matsuda, for being on my committee and providing his inputs and valuable feedback about the work. I thank my parents, for always being my source of inspiration and motivation. They always encouraged me to pursue my desires in an extraordinary way. I thank my sister Ashmeet Kaur for her all time support and help. I thank all my friends for their advice and cooperation at every moment. Finally, I am thankful to God, for granting me the capability, skills and opportunity that made this possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER

1

INTRODUCTION

1.1 Definitions

1.1.1 Peer Assessment

"Peer assessment is an arrangement for learners to consider and specify the level, value, or quality of a product or performance of other equal-status learner" [1]. Peer assessment is basically a collaborative approach to evaluating the performance of individuals within a group, in which students assess each other's work and provide feedback. It is becoming increasingly popular in educational settings as it has been found to have numerous benefits, including enhancing critical thinking, fostering teamwork and communication skills, and improving motivation and engagement [1; 2; 3]. It involves assessing one's work by one's peers, aiming to improve learning outcomes and encourage a deeper understanding of the subject matter. Peer assessment can be used in various educational settings, including schools, colleges, and universities.

One of the main advantages of peer assessment is that it allows students to receive feedback from their peers, which can be more relevant and helpful than feedback from a teacher or instructor [4]. This is because peers have a better understanding of the challenges

and difficulties that their classmates face and can offer constructive criticism that is based on their own experiences. Peer assessment can also be an effective way to develop students' self-assessment skills, helping them to gain a better understanding of the criteria used to evaluate academic work and develop a more nuanced and sophisticated approach to their own learning [2].

To ensure the effectiveness of peer assessment, it is important to establish clear criteria and guidelines for the assessment process, provide adequate training and support to students, and ensure that the assessment is fair and unbiased. Overall, peer assessment can be a valuable tool for promoting active and collaborative learning and can help to prepare students for the demands of the real world, where teamwork and communication skills are essential [3].

1.1.2 Expertiza

Expertiza is an open-source tool for collaborative and constructive feedback, allowing students and professionals to exchange critiques on academic papers, research projects, and other written work. Expertiza offers a range of features and customization options, including the ability to create customized rubrics and evaluation criteria, anonymous reviews, self-review, and peer moderation [15]. It has been widely used in computer science and engineering courses, as well as in other disciplines, to support the development of writing skills and critical thinking abilities.

Expertiza has also been adopted by professional organizations and research communities as a tool for peer review and collaboration. For example, the Association for Computing Machinery (ACM) uses Expertiza for peer-reviewing submissions to its conferences and journals [17].

The main challenge in Expertiza is the difficulty of ensuring that students engage in meaningful and constructive feedback rather than simply going through the motions to fulfill a requirement.[18]

1.1.3 Task of Crowd Labeling

Crowd labeling, also known as crowd tagging, refers to the process of outsourcing the annotation or tagging of data to a large group of people/students, usually via an online platform. For our study, we are using Expertiza as an online platform for crowd-labeling and data annotation. This labeling method has become increasingly popular in recent

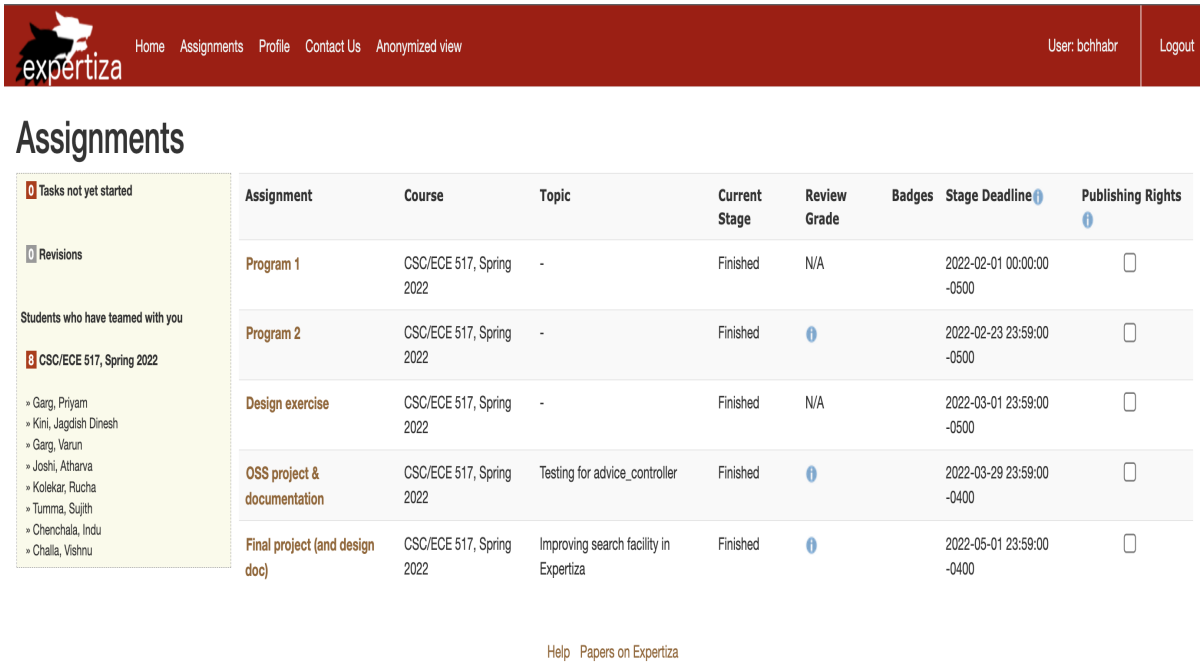

Figure 1.1: Screenshot of Expertiza platform

years due to its efficiency, cost-effectiveness, and ability to handle large amounts of data. Crowd labeling is used in various fields, such as natural language processing, computer vision, and machine learning. It involves presenting workers with data, such as images or text, and asking them to label or tag it as belonging to specific categories or having specific attributes. Labeling is increasingly being used for various types of data, including textual data. Textual data can be annotated for various tasks, such as sentiment analysis, entity recognition, and text classification. Crowd labeling enables large amounts of textual data to be annotated quickly and accurately as the work is distributed among many workers, reducing the likelihood of bias. One of the major advantages of crowd labeling is that it allows for a large amount of data to be annotated quickly and accurately. This is because the work is distributed among many workers, each of whom may have different expertise or perspectives, thereby reducing the chances of bias in the labeling process. Additionally, crowd labeling can be more cost-effective than hiring and training in-house annotators.

Online platforms such as Amazon Mechanical Turk, CrowdFlower, and Figure Eight provide access to a large pool of workers and tools for managing and validating the annotation process [8]. These platforms have been used for tasks such as sentiment analysis [7], entity recognition, and text classification [5].

However, crowd labeling also has some drawbacks. For example, there may be issues

with quality control, as workers may have different levels of expertise or motivation. The reliability and consistency of the crowd remain a challenge in crowd labeling, especially for textual data. Researchers have explored various techniques to ensure the quality of the annotations, such as annotation agreement measures and iterative learning techniques [5; 6].

1.1.4 Overview of Strategies for Quality Control of Crowd Labeling

Crowd labeling, or crowd annotation, is a popular and cost-effective way to generate labeled data for machine learning tasks. However, ensuring the quality of the labeled data is essential for machine learning tasks, but it can be a challenging task, especially when the crowd is large and diverse. One strategy to improve the quality of the labeled data is to implement a fast tagging quality control strategy that identifies workers who tag too fast, i.e., workers who assign labels too quickly without adequate consideration. These workers may be either uninterested or unskilled, and their labels are more likely to be low-quality, inconsistent, or noisy. Fast tagging quality control strategies aim to identify and filter out low-quality labels, allowing for high-quality labels to be prioritized for use in training models based on identifying students who tag "too quickly" and assigning tags so fast that they could not have given adequate time to think about the reviews and consider whether the review supports the tag category or opposes it.

This approach to identifying workers who tag too fast is to monitor the time taken by each worker to assign labels. Workers who assign labels too quickly may be identified as potential low-quality contributors. Another approach is to use a minimum time threshold to ensure that workers have taken enough time to consider their assigned labels.

Studies have shown that such quality control strategies that identify workers who tag too fast effectively improve the quality of crowd-labeled data. For example, a study proposed a quality control strategy that uses a minimum time threshold for labeling tasks. They found that this approach helped identify low-quality workers and significantly improved the quality of labeled data.[9]

Similarly, in a study, the authors proposed a quality control strategy that monitored the time taken by workers to assign labels and identified workers who completed labeling tasks too quickly. They found that this approach helped identify low-quality workers and reduced the number of inconsistent labels. [10; 41]

Another strategy for improving the quality of crowd labeling is to identify workers who provide inconsistent labels, i.e., who assign labels in disagreement with other team

members. One approach to identifying workers who provide inconsistent labels is to use an inter-rater reliability (IRR) measure. IRR measures the degree of agreement among multiple raters on the same labeling task. Workers who provide labels that deviate significantly from the consensus of other workers may be identified as potential low-quality contributors.

IRR can be measured in various ways, such as Fleiss' kappa, Cohen's kappa, and Krippendorff's alpha. Once the IRR is calculated, a threshold can be set to identify workers whose labeling performance falls below a certain level. Once the low-performing raters are identified, they can be given additional training, feedback or removed from the tagging task. In addition, consensus-based methods can be employed to reconcile conflicting tagging and obtain a final label. Workers who fall below the threshold can be excluded from further labeling tasks or asked to undergo additional training.

Studies have shown that quality control strategies that identify workers who provide inconsistent labels effectively improve the quality of crowd-labeled data. For example, in a study, the authors proposed a quality control strategy that used an IRR measure to identify low-quality workers in a text classification task. They found that this approach helped identify low-quality workers and significantly improved the quality of labeled data.[12]

Similarly, in another study the authors proposed a quality control strategy that used a weighted IRR measure to identify low-quality workers in a video annotation task. They found that this approach helped identify low-quality workers and reduced the number of inconsistent labels.[13] Another study that utilized this quality control strategy is "Crowdsourcing Annotation of Clinical Texts: Measuring Annotator Engagement and Agreement" by Pradhan [11]. The study used IRR to identify raters with the low agreement and found that providing feedback and additional training improved their performance. Another study, "Crowdsourcing with Expert Feedback and Its Application to Medical Image Annotation" used consensus-based methods to reconcile conflicting annotations and achieved high accuracy in the final labels.

Krippendorff's alpha is a commonly used measure for assessing the inter-rater agreement in crowdsourcing tasks. It has been shown to be effective in identifying low-quality raters and improving the overall quality of the tagging.

In Krippendorff's alpha, the agreement between each rater and the majority of other raters is measured. A rater whose tags have a low agreement with the majority of other raters is considered a low-quality rater, and their tags are filtered out. The remaining tags are used to compute the final consensus tagging.

Another quality control strategy is to identify unreliable workers who are tagging in a particular pattern, such as yes-no-yes-no-yes-no, indicating that they are not paying

attention to the task and are simply choosing answers randomly. Identifying such unreliable workers can help improve the overall quality of crowd labels by removing their contributions from the dataset.

A study by Lease, Matthew [14; 40] proposed a quality control strategy for crowd labeling, which involves identifying unreliable workers who are tagging in a particular pattern. The study analyzed the labeling behavior of crowd workers on a textual classification task and identified unreliable workers based on their tagging patterns. The study showed that by removing the contributions of these unreliable workers, the overall quality of crowd labels improved significantly.

Labeler calibration is another quality control strategy for crowd-labeling textual data. It involves ensuring that the labelers assigned to a particular task are reliable and consistent in their labeling. This is accomplished by comparing the labels assigned by each labeler to the predictions made by an algorithm based on Yulin's prediction[32]. According to Yulin's prediction [32], if a tagger repeatedly disagrees with high-probability tags, they are probably wrong. Therefore, if a labeler's labels consistently agree with the algorithm's predictions, then the labeler is considered reliable.

The labeler calibration process involves several steps. First, a group of labelers is selected to complete a labeling task. Next, the labels assigned by each labeler are compared to the predictions made by an algorithm. If a labeler's labels consistently agree with the algorithm's predictions, then the labeler is considered reliable and is given a high rating. If a labeler's labels consistently disagree with the algorithm's predictions, then the labeler is considered unreliable and is given a low rating. The ratings are used to determine which labelers are assigned to future labeling tasks.

Labeler calibration is important because it helps ensure the accuracy and consistency of the labeled data. Inaccurate or inconsistent labels can lead to incorrect conclusions and analysis. By using a reliable and consistent group of labelers, researchers can increase the reliability and validity of their findings.

The study used in our project used labeler calibration to improve the quality of labeled data was conducted by Yulin. The study used an active-learning approach to reduce the amount of labeling effort required while maintaining reliable training data. The results of the study [32] showed that they were able to cut the amount of labeling effort required by half without a loss of reliable training data. This approach can be particularly useful when labeling large datasets that would otherwise require a significant amount of time and effort to label.[32]

Another strategy used in this research is to explore the relationship between the IRR val-

ues for tags and the level of agreement among crowd workers within a team. We hypothesize that a high degree of consensus signifies a strong level of reliability for the corresponding tag. This consensus among crowd labelers can serve as a crucial strategy for identifying instances where a student's response deviates from the majority consensus, thereby highlighting potential areas of concern in the labeling process.

By examining the interplay between IRR values and tagging consensus, our research contributes to enhancing the quality control mechanisms in crowd labeling. The insights gained from this study can inform the development of more robust and reliable systems for crowd-generated label data, facilitating improved decision-making and downstream applications in various domains relying on crowd-sourced data.

In the subsequent sections, we will present our methodology, describe the data collection process, and discuss the results and implications of our findings. By shedding light on the importance of our quality control strategies for tag consensus, we strive to offer practical recommendations and guidelines for improving the quality and reliability of crowd-labeling efforts, ultimately advancing the field of crowd-based data annotation.

1.2 Problem and Motivation

Crowd labeling has emerged as a popular approach to collecting large datasets for various machine-learning tasks. However, the reliability of the labels provided by a diverse crowd can vary significantly, making it challenging to build accurate models. Quality control of crowd labeling is, therefore, crucial to ensure that the collected data is reliable and can be used to build effective models.

The problem that our study addresses is the lack of reliable and high-quality data in crowdsourcing tasks. Crowdsourcing has become an increasingly popular approach for data labeling, annotation, and analysis due to its low cost, scalability, and flexibility. However, ensuring the quality of the labeled data remains a major challenge, as the accuracy and consistency of the labels can vary widely among the crowd workers.

One area where crowd labeling has gained significant attention is peer-assessment-based educational methods. The rise of these methods has resulted in the need for effective quality control of crowd labeling. These methods rely on student reviews of their peers' projects and comments, which are then used to provide suggestions for improvement. However, the effectiveness of these methods depends on the quality of the comments and tags provided by the students. In this context, the quality control of crowd labeling is a

significant challenge that needs to be addressed.

The main challenge in ensuring the quality of crowd labeling in educational settings is that the taggers are not trained professionals, and their tagging accuracy can vary significantly. Therefore, it is essential to develop effective quality control mechanisms to ensure the reliability of the tags provided by students.

The research presented in this paper aims to tackle this challenge by developing some natural language processing approaches to evaluate the quality of student tags. The effectiveness of these approaches relies on the quality of the training data, which consists of comments labeled by students as containing or not containing certain characteristics. Thus, it is critical to validate the quality of the student tagging to ensure the accuracy of the training data.

To address this, several strategies have been implemented, including identifying students who tag too fast, excluding tags where different students disagree, identifying patterns in tags, and using labeler calibration to compare student tags with predicted tags from the training data. These strategies will help assign reliability metrics to each tag and tagger, providing researchers with filtered and larger datasets that include only the most reliable tagged data.

The motivation behind this research is to improve the quality of crowd labeling in educational settings, which will lead to better feedback and suggestions for improvement for students. The strategies developed in this research will provide reliability metrics for each tag and tagger, enabling researchers to filter the most reliable tagged data or use larger datasets that also include less reliable tags. This will ultimately help improve the quality of crowd labeling in educational settings, leading to better feedback and suggestions for improvement for students and more effective peer-assessment-based educational methods.

Our study aims to address a pressing problem in crowdsourcing research and provide a practical solution for improving the quality of the labeled data. The potential applications of our approach are numerous and can benefit researchers in a wide range of fields, from natural language processing and computer vision to social media analysis and beyond.

CHAPTER

2

RELATED WORK

Crowd labeling has become a popular way to collect data for machine learning models. However, the quality of the labeled data is critical to the success of these models. Many studies have focused on improving the accuracy and reliability of labels assigned by the crowd, including students. In recent years, several studies have been conducted on quality control of crowd labeling. Related work includes previous research on methods for evaluating and improving the accuracy of crowd-sourced annotations

One of the most common approaches used in such studies is an inter-rater agreement, where multiple raters are assigned to label the same data, and the degree of agreement between them is calculated. In one study, the authors used inter-rater agreement as a quality control measure for the labels assigned by a crowd of Amazon Mechanical Turk workers. The authors found that filtering out low agreement labels improved the performance of machine learning models trained on the data. In the field of natural language processing, a study proposed a method for improving the quality of crowd-sourced sentiment analysis tasks by filtering out low-quality annotations based on inter-annotator agreement and label entropy. The authors found that their method improved the accuracy of the sentiment analysis task by up to 14 percent[33]. Additionally, several platforms exist for managing crowd-sourced annotations, including Amazon Mechanical Turk and CrowdFlower, which

provide built-in quality control mechanisms such as worker reputation systems and gold standard tasks to evaluate worker accuracy [34].

Inter-rater reliability is a widely used strategy for evaluating the consistency of the labels assigned by multiple labelers. Krippendorff's alpha is a commonly used measure of inter-rater reliability that takes into account both the observed agreement among raters and the expected agreement by chance. Several studies have used Krippendorff's alpha to evaluate the reliability of crowd-labeled data. In a study by Snow et al. (2008)[38], the authors used Krippendorff's alpha to evaluate the inter-rater reliability of labels assigned to a set of sentences by a crowd of Amazon Mechanical Turk workers. The authors found that Krippendorff's alpha was a useful measure of inter-rater reliability and could be used to identify low-quality labels.

Another study by Kulesza [39] used Krippendorff's alpha to evaluate the reliability of labels assigned to a set of user reviews by a crowd of workers. The authors found that Krippendorff's alpha was effective in identifying low-quality labels and could be used to improve the quality of the crowd-labeled data. In addition to evaluating the reliability of crowd-labeled data, Krippendorff's alpha can also be used to compare the quality of labels assigned by different labelers. In another study, the authors used Krippendorff's alpha to compare the quality of labels assigned by experts and non-experts to a set of medical records. The authors found that Krippendorff's alpha was an effective measure for evaluating the quality of labels assigned by both expert and non-expert labelers.

However, it should be noted that Krippendorff's alpha has some limitations and assumptions. For example, it assumes that the labels are independent and that the agreement between labelers is not due to chance. In addition, Krippendorff's alpha may not be appropriate for all types of data and labeling tasks. Overall, Krippendorff's alpha is a useful measure for evaluating the reliability and quality of crowd-labeled data, but it should be used in conjunction with other quality control measures to ensure the accuracy and consistency of the labels.

Another approach is to identify fast labelers and remove their labels from the dataset. Fast tagging, also known as speed-based filtering, is a strategy to identify and remove labels assigned by fast labelers from the dataset. The idea behind this strategy is that fast labelers may not pay enough attention to the task and may produce lower-quality labels. Several studies have explored the effectiveness of fast tagging in improving the quality of crowd-labeled data. In a study, Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions [35], the authors developed a method to identify fast labelers and exclude their labels from the dataset. The authors used

a threshold based on the median time taken to label each item and found that excluding the labels assigned by fast labelers improved the accuracy of the labels. In this study, the authors found that fast labelers were less accurate than slower labelers, and hence they used this information to develop the method to identify fast labelers and exclude their labels from the dataset. Another study investigated the effects of fast tagging on the quality of labels assigned by a crowd of Amazon Mechanical Turk workers. The authors found that removing labels assigned by fast labelers improved the accuracy of the labels and reduced the variability between different labelers [36]. In a study by Liu [14; 37], the author found that fast tagging resulted in a loss of high-quality labels and recommended using a hybrid approach that combines fast tagging with other quality control measures. Overall, fast tagging is a useful strategy for improving the quality of crowd-labeled data, but it should be used in combination with other quality control measures to ensure that high-quality labels are not excluded from the dataset.

Another approach involves using gold-standard data, where a subset of the data is pre-labeled by experts, and the performance of the crowd is evaluated based on their ability to match the expert labels. Some of these studies have focused on improving the quality of crowd labeling by using active learning techniques. Active learning involves iteratively selecting samples that are uncertain or difficult to classify and requesting labels from the crowd for those samples. This approach can improve the accuracy of the labels while reducing the overall number of labels required.

The pattern detection strategy is another strategy for quality control of crowd labeling that identifies systematic errors or biases in the labeling process based on patterns in the tag dataset and has been widely used in quality control for crowd-labeled data. This strategy aims to detect patterns such as all tags set to yes, all tags set to no, alternating yes and no, or repeated sequences such as “yes, yes, no, yes, yes, no, yes, yes, no.” These patterns can be indicative of labelers who are not paying attention to the content or who are intentionally labeling in a biased manner.

One such algorithm used to detect patterns in the data is the Progressive Timelist-Based Verification (PTV) algorithm. PTV is a tree-based method that can detect periodic patterns in the data, such as repeated sequences of tags or variations in the frequency of different tags. This algorithm has been used in a number of studies to identify patterns in crowd-labeled data, including text classification and image tagging.

For example, in a study by Wang, PTV was used to detect patterns in a dataset of user-generated image tags. The authors found that PTV was able to identify systematic errors in the tagging process, such as taggers who consistently assigned incorrect or irrelevant tags.

They also found that PTV was able to identify taggers who were consistently inconsistent in their labeling behavior, indicating a lack of attention to the task.

Another study used a similar approach to detect patterns in a dataset of product reviews. The authors used a combination of PTV and clustering algorithms to identify groups of taggers with similar labeling behavior, allowing them to identify taggers who were consistently biased or unreliable. The paper "Mining Periodic Patterns in Sequence Data" by Agrawal [42], is another and the most effective study related to our project. It proposes a method for mining periodic patterns in sequence data. The authors present an algorithm called Progressive Timelist-Based Verification (PTV) that can efficiently mine periodic patterns in large sequence databases. The PTV algorithm works by first generating a set of candidate periodic patterns, and then using a verification step to prune the set of candidates that do not meet the periodicity constraints. The main findings of the paper are as follows: The PTV algorithm is an efficient method for mining periodic patterns in large sequence databases. The authors demonstrate the effectiveness of their method using several real-world datasets, including stock prices and earthquake data. The PTV algorithm can be used to discover interesting periodic patterns in sequence data, such as daily or weekly patterns in stock prices, or seasonal patterns in earthquake data. The authors also propose several extensions to the PTV algorithm, including a method for mining multi-level periodic patterns and a method for mining patterns with gaps.

The authors evaluate the effectiveness of their method using several real-world datasets, including stock prices, weather data, and earthquake data. They demonstrate that the PTV algorithm is able to discover interesting periodic patterns in these datasets, such as daily or weekly patterns in stock prices, seasonal patterns in weather data, and periodic patterns in earthquake data. They also compare the performance of the PTV algorithm with several other methods for mining periodic patterns and show that the PTV algorithm outperforms these methods in terms of both efficiency and effectiveness. This paper provides a useful method for mining periodic patterns in sequence data, which has important applications in fields such as finance, biology, and environmental science. The PTV algorithm is a powerful and efficient method for discovering periodic patterns, and the extensions proposed by the authors make it even more versatile and useful. The paper also highlights the importance of periodic pattern mining in understanding and predicting complex systems and opens up new avenues for future research in this area. Overall, pattern detection strategies such as PTV can be an effective tool for identifying systematic errors or biases in crowd-labeled data and can help improve the accuracy and consistency of the labels.

Overall, previous research has demonstrated the importance of quality control in crowd-

sourced annotation tasks, and various methods have been proposed and tested to improve annotation accuracy. The strategies outlined in this paper for validating the quality of student tagging in peer-assessment-based educational methods add to the existing literature on quality control of crowd labeling. These related works demonstrate the importance of quality control in crowd labeling, and the various approaches that can be used to ensure the reliability and accuracy of the labeled data.

CHAPTER

3

STUDY OBJECTIVE AND APPROACH

In this chapter, we provide our study objective and approach used for the quality control of crowd labeling for peer assessment-based educational methods. We specify what all metrics are used to identify reliable taggers for tagging the review categories. This chapter covers what our study is trying to achieve and summarizes the approach used to achieve the results.

3.1 Study objective

This study aims to improve the quality of student-generated labels (tags) in peer assessment-based educational methods. The objective is to establish a framework that ensures the reliability and accuracy of these tags, enhancing the overall effectiveness and credibility of student-generated labels. By addressing this objective, the research seeks to contribute to the advancement of automated grading systems, offering promising prospects for streamlining assessment processes in educational and professional settings.

To achieve this objective, several strategies and techniques will be implemented and evaluated. These include excluding tags and taggers with low inter-rater reliability, identify-

ing fast taggers, detecting patterns between tags, and utilizing labeler calibration. These strategies aim to assign reliability metrics to each tag and tagger, enabling the establishment of thresholds for including tags in the dataset and grading students based on their reliability.

Furthermore, the study aims to examine the effectiveness of these strategies in improving the quality of training data for machine learning models used in peer assessment-based educational methods. By filtering out unreliable tags and taggers, the resulting datasets will be more reliable, accurate, and suitable for training machine learning models. This will enhance the performance and effectiveness of the machine learning models in supporting peer assessment processes.

Additionally, this research aims to reveal significant parallels and consensus between the metrics generated by our implemented strategies and the manual grading conducted by professors. The objective is to provide empirical evidence supporting the consideration of our approach as a viable alternative to manual grading in crowd-labeling contexts. The compelling results obtained from the comparative analysis reinforce the potential of our methodology in enhancing efficiency, accuracy, and cost-effectiveness in various domains relying on crowd-sourced data.

By adopting our study's methodology, educational and professional domains can benefit from the advancements in automated grading systems. The findings contribute to the growing body of research on automated assessment processes, presenting promising prospects for streamlining evaluation and feedback mechanisms. Our research serves as a stepping stone towards leveraging technology and machine learning algorithms to improve the speed and consistency of grading tasks.

Furthermore, the implications of this research extend beyond the educational realm. The successful implementation of our strategies for quality control in crowd labeling opens up opportunities for other domains reliant on crowd-sourced data, such as market research, sentiment analysis, and data annotation tasks. Embracing our approach has the potential to revolutionize these domains by ensuring reliable and accurate results while reducing time and resource expenditures.

This study's objective is to improve the quality of student-generated labels in peer assessment-based educational methods through the implementation of various strategies. These strategies aim to ensure the reliability and accuracy of tags and taggers, resulting in enhanced training data for machine learning models and improved assessment processes. The research findings reveal significant parallels and consensus between our strategies' metrics and manual grading, supporting the consideration of our approach as a viable alternative to manual grading in crowd-labeling contexts. By adopting our methodology,

various domains relying on crowd-sourced data can enhance their efficiency, accuracy, and cost-effectiveness. This research contributes to the advancement of automated grading systems, offering promising prospects for streamlining assessment processes in educational and professional settings.

3.2 Study Approach

The study focuses on quality control of crowd labeling in peer assessment-based educational methods. The first step is to identify fast taggers. It is important to identify taggers who assign labels too quickly without proper consideration. These taggers may be either uninterested or unskilled, and their labels may be of low quality, inconsistent, or noisy. This is achieved by identifying students who assign tags too quickly without giving adequate time to think about the reviews and consider whether the review supports the tag category or opposes it.

The second step is to identify inter-rater reliability between students of the same team. Students with low inter-rater reliability signify a disagreement between the team members, and hence some of the team members could be considered unreliable taggers based on their probabilistic disagreement value.

The third step is to detect patterns between tags. It is important to identify taggers who are not paying attention to the task and are tagging in a repetitive pattern, such as yes-no-yes-no-yes-no. Such taggers could be considered unreliable and may negatively affect the overall quality of the dataset. Detecting these unreliable taggers and removing their contributions from the dataset can help improve the overall quality of crowd labels.

Another step in the study approach involves the utilization of labeler calibration to compare the tags assigned by students with the predicted tags derived from the training data. This step aims to assess the alignment between the students' tag assignments and the expected tags based on the collected training data. By comparing these two sets of tags, we can evaluate the accuracy and reliability of the student-assigned labels.

Additionally, the quality of the labeled data is further evaluated through the calculation of inter-rater reliability (IRR) values for each tag. IRR is a statistical measure that quantifies the level of agreement among multiple raters, which, in this case, refers to the crowd workers involved in the tagging process. The IRR values are computed using various metrics, and for this particular study, the mode metric is employed. The mode metric identifies the most frequently occurring tag label for a given tag prompt.

Once the IRR values for each tag are obtained, a consensus-based approach is employed to determine the consideration of each tag. The study examines the level of agreement among the crowd workers for a specific tag. If we have a high degree of agreement, such as for a value of 1, it suggests a strong consensus among the crowd workers, indicating a reliable and consistent tag value to be 1.

These strategies aim to assign reliability metrics to each tag and tagger, allowing for the establishment of thresholds for including tags in the dataset and grading students based on their reliability. The ultimate goal is to improve the accuracy and dependability of peer-reviewed comments by enhancing the quality of tags assigned by a group of students. This is important for training machine learning models to recognize effective peer reviews and tagger reliability automatically. The results of the strategies used in the project will be examined and provided to researchers for further evaluation.

Finally, we provided researchers with a comprehensive report and results of our findings, including the IRR values, the table of tags and their reliability, and recommendations for best practices in crowdsourcing quality control. Our study approach aimed to be transparent, reproducible, and adaptable to different types of crowdsourcing tasks and data samples.

CHAPTER

4

EXPERIMENTAL SETUP AND STRATEGIES

4.1 Experimental Setup, Math, and Libraries Used

4.1.1 Data Collection

Data collection is a crucial step in any research project, and in this study, we are collecting data from Expertiza. Expertiza is an open-source tool for collaborative and constructive feedback used by students, professors, and TAs. The platform allows for assignment submission, peer assessment, grading, and feedback.

To begin with, the data collection process involved accessing Expertiza and reviewing the assignments submitted by the teams. It is important to note that the assignments are created in Expertiza by the professor and TAs. Students are grouped into a number of teams for each assignment. These teams submit assignments, and other teams give reviews of these submissions. The team that receives these reviews and comments then provides feedback in the form of tags.

Tags, also known as labels, are a specific form of feedback that teams provide in response to peer reviews of their submitted assignments. Tags are a form of structured feedback that allows team members to provide concise and descriptive feedback on specific aspects of

the reviewed assignment.

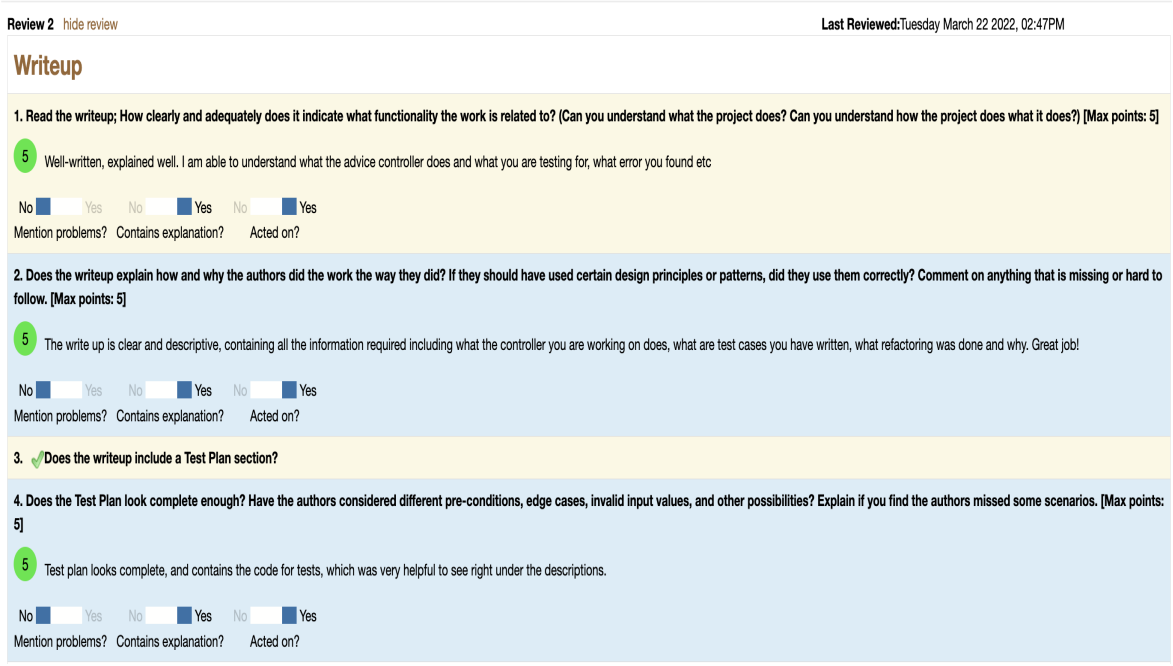

Figure 4.1: Demonstration of a tag and its categories

Tags allow for the quick and efficient communication of feedback and can help to highlight common issues or strengths in the reviewed assignments and provide consistency in feedback. They also allow for the easy analysis and categorization of feedback data, which can be useful for identifying trends and patterns in the feedback provided by different teams. Tags are used to provide structured feedback on specific aspects of an assignment. Tags are pre-defined keywords or phrases that represent common issues, strengths, or suggestions for improvement in the assignment.

The next step in data collection involved reviewing the tags provided by the teams. These tags represent the feedback provided by the team, and they are the dataset used in this research. The tags were analyzed and categorized according to the specific assignment they were associated with. This categorization was important in identifying common themes and patterns in the feedback provided.

It is worth noting that Expertiza offers several features and customization options, including the ability to create customized rubrics and evaluation criteria, anonymous reviews, self-review, and peer moderation. These features were considered in the data

collection process to ensure that the feedback provided was accurate and reliable.

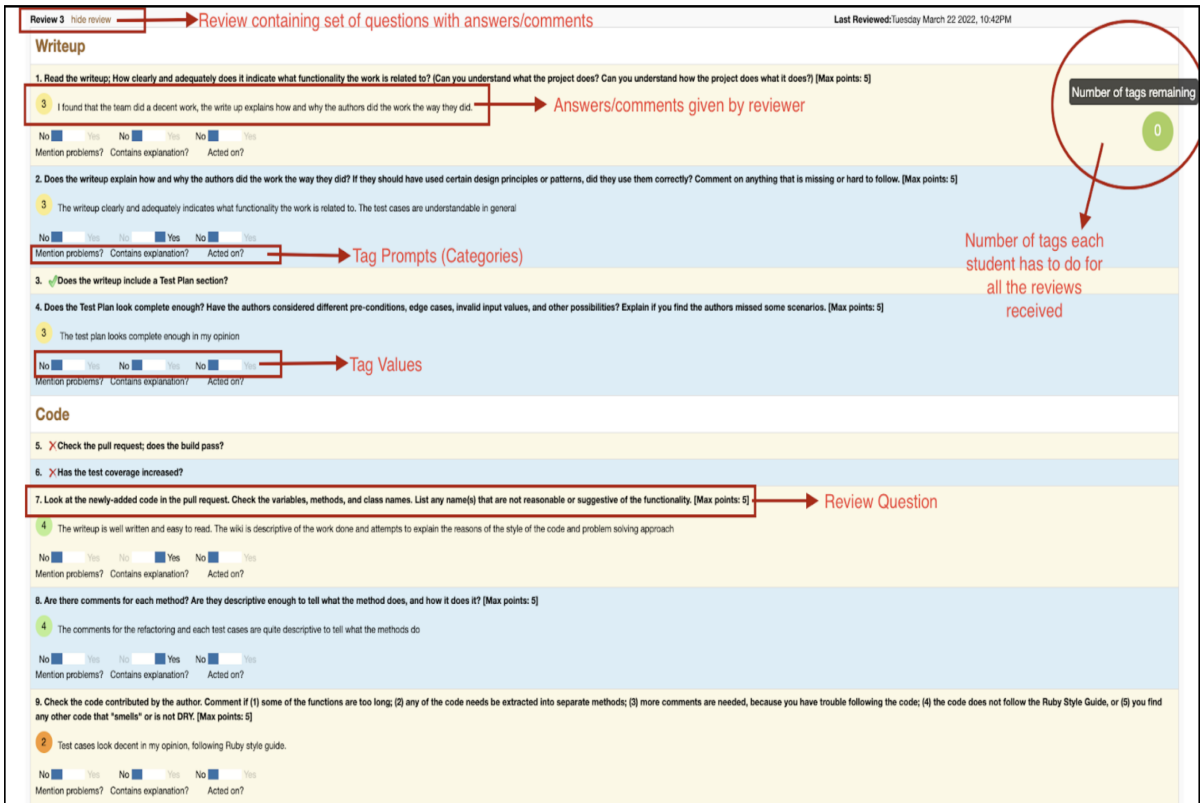

Figure 4.2: Demonstration of a tags

In this study, a dataset comprising peer-reviewed comments is collected from the study participants. The purpose of collecting this dataset is to serve as a basis for evaluating the proposed strategies for quality control. Each comment in the dataset is tagged by multiple participants, enabling the assessment of inter-rater reliability among the participants

The process of data collection is conducted through the utilization of an online platform called Expertiza, as mentioned earlier. This platform facilitates the collection of peer reviews and the assignment of tags by a group of students. The participants have the opportunity to submit their reviews and assign corresponding tags using the platform. The reviews provided by the students are in text format, while the tags associated with these reviews are represented using sliders.

To ensure the high quality of the collected data, clear instructions are provided to the students. These instructions aim to guide the students in writing effective reviews and

assigning accurate tags. The guidelines encompass various aspects, including the provision of constructive feedback, the appropriate assignment of tags based on the content of the reviews, and the avoidance of bias in their reviews and tags.

By providing these comprehensive instructions, the study aims to foster a standardized and reliable data collection process, enhancing the overall quality of the collected dataset. This meticulous approach ensures that the data utilized for evaluating the strategies is accurate, consistent, and reflective of the participants' insights and assessments.

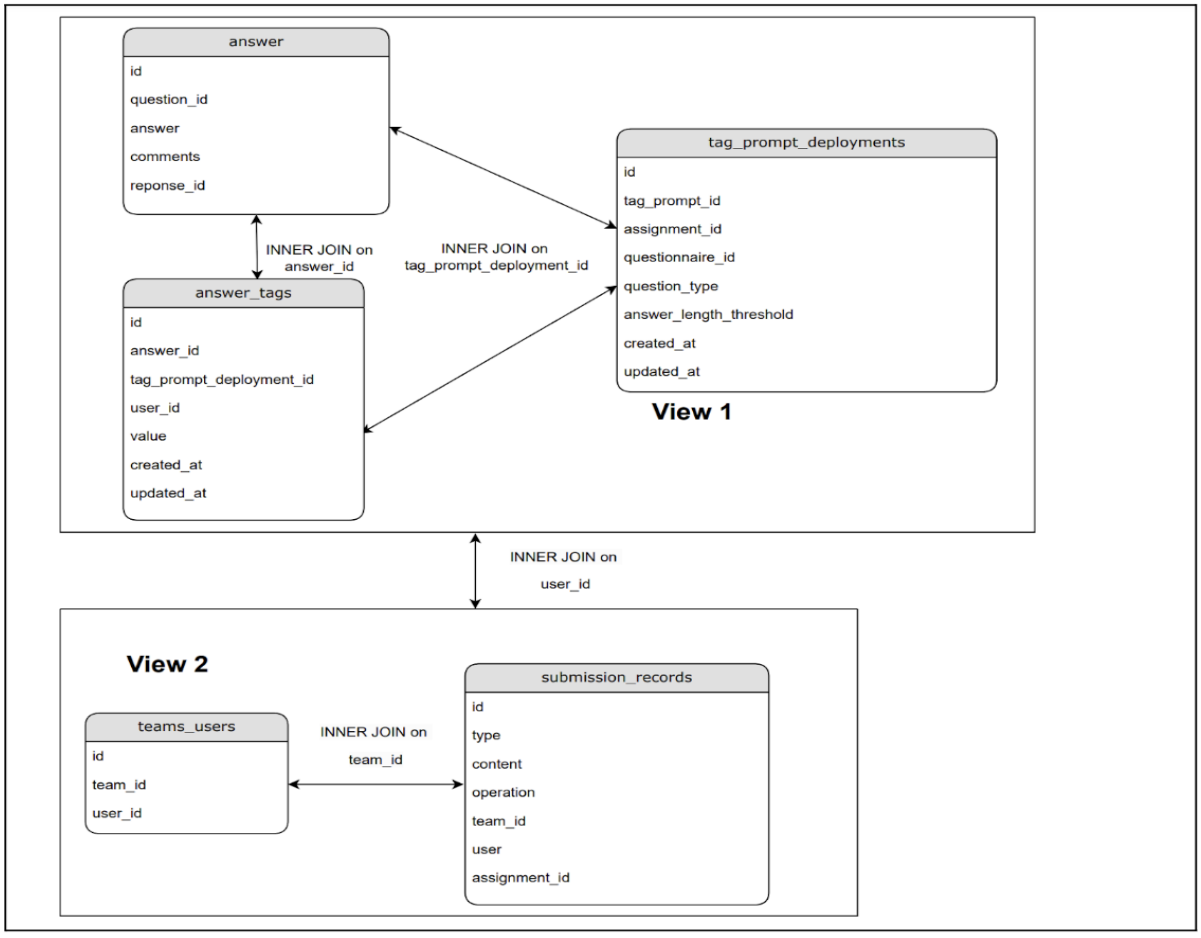

Figure 4.3: Demonstration of Expertiza Database Tables Used

We conducted data collection from the Expertiza database. The database serves as a valuable source of information for our research, and we utilized data from 5 specific tables out of the total 80 tables available. These tables include answers, answer_tags,

tag_prompt_deployments, submission_records, and team_users.

To extract relevant and comprehensive data, we performed inner joins on these 5 tables. The process of joining these tables is depicted in the picture above, illustrating the relationships and connections between them. Specifically, we executed joins between the _tags, and tag_prompt_deployments tables to create View 1. This view is based on the tag_prompt_deployment_id and _id, allowing us to consolidate essential data.

Additionally, we conducted a join between the submission_records and team_users tables on the teams_users field, resulting in View 2. This view contains essential information about user teams, which is crucial for our data analysis.

In order to obtain a holistic dataset with all the required fields to calculate the desired results for each strategy, we combined View 1 and View 2. This final step involved a join on the user_id field, bringing together all the relevant information from both views.

The database connectivity and joins are visually represented in the image provided above, offering a clear overview of the data collection process. This comprehensive approach enables us to access the necessary data for evaluating the effectiveness of each quality control strategy in crowd labeling.

Through these carefully executed data collection methods, we aim to gain valuable insights into the reliability and accuracy of crowd-generated labels in peer assessment-based educational settings. The data acquired from the Expertiza database serves as a foundational resource for our research, supporting our efforts to enhance the quality of crowd labeling in various applications.

4.2 Quality Control Strategies

This study aims to examine the validity and reliability of reviews and tags obtained from peer assessment activities, which typically involve the receipt of a score that may be included as part of the final grade. To examine the validity and reliability of taggers and tags, we focus on the implementation of four strategies, the results of which will be provided to researchers for further evaluation.

4.2.1 Strategy 1: Fast Tagging

The fast tagging quality control strategy is a valuable tool for ensuring the accuracy and consistency of crowd labeling for textual data. This strategy is particularly important when dealing with large volumes of data, where it can be challenging to ensure the quality of all

contributions. By identifying taggers who tag too quickly and may be providing low-quality labels, this strategy helps to filter out unreliable contributions, allowing high-quality labels to be prioritized for training machine learning models.

One of the key advantages of this strategy is its ability to detect potential low-quality contributors based on the speed at which they assign labels. Taggers who assign labels too quickly may not have given adequate time to consider the review or the tag category, leading to inconsistent or noisy labels. Identifying these taggers will help to improve the overall quality of crowd labeling for textual data.

We use a minimum time threshold to identify taggers who tag too fast. By setting a minimum time threshold, taggers who didn't take enough time to consider their assigned labels and hence tagged quite quickly could be considered unreliable taggers, which helps in reducing the risk of low-quality or inconsistent labels. This approach can be particularly effective when combined with other quality control strategies, such as pattern detection and inter-rater reliability analysis.

In addition to improving the quality of labeled data, our strategy can also help to increase the efficiency of crowd labeling for textual data. By prioritizing high-quality labels, machine learning models can be trained more effectively, reducing the need for manual review and improving the speed and accuracy of the overall labeling process. Another benefit of using a fast tagging strategy for quality control is that it can help to reduce the workload of manual review by flagging potentially low-quality labels for manual inspection. Instead of reviewing every single label, the manual review process can focus on those labels that are flagged by the fast tagging strategy, which can save time and effort.

To implement this strategy effectively, it is important to establish clear guidelines and standards for tag assignment. Taggers are trained to ensure they understand the tag categories and can assign them accurately and consistently. Additionally, taggers are encouraged to take their time when assigning labels rather than rushing through the process to meet a quota.

It's worth noting that while fast tagging quality control can be an effective strategy for identifying low-quality contributors, it is not always foolproof. Some taggers may be able to assign labels quickly while still producing high-quality labels, and conversely, some workers may take a long time to assign labels but still produce low-quality labels. As with any quality control strategy, it's important to balance the need for high-quality labels with avoiding excluding potentially valuable contributors based on overly strict quality control criteria. Hence to ensure the correctness and effectiveness of quality control strategies, it is important to combine some of the strategies to get the best results. Our Fast tagging strategy

can also be combined with other quality control strategies, such as pattern detection and inter-rater reliability checks. For example, a tagger who tags too quickly may also be flagged as unreliable in inter-rater reliability checks, indicating that their labeling may be of lower quality than other workers. Using multiple quality control strategies can help to ensure that the final labeled dataset is of high quality and can be used for training machine learning models.

We are implementing this strategy by using the timestamp information of each tag a worker assigns. In this approach, the answer tags data of each student user_id for each assignment are sorted in ascending order of their tag timestamp. Then, the difference between consecutive timestamps of a student is calculated and added to calculate the total time taken by that student to complete the assignment. By averaging the total time taken over all tags for each student, we measure the speed of tagging.

Once the average tagging speeds of all students have been calculated, we further refine our quality control approach by applying a logarithmic transformation to the average values. The logarithmic transformation can help to compress the distribution of the average values, making it easier to identify students who are tagging too quickly. This transformation is particularly useful when dealing with large datasets with many students, where the distribution of average values may be highly skewed.

Using the logarithmic average instead of the plain average of tag times helped us to mitigate the effect of long breaks between tagging sessions. If a student takes a long break, it will artificially inflate the plain average of their tag times, making it harder to distinguish between students who are truly fast taggers and those who took a break. By taking the logarithm of the average tag time, the effect of long breaks is scaled down, and the resulting score is a better reflection of the student's actual tagging speed. This can help to identify fast taggers more accurately and remove their low-quality tags from the dataset, leading to improved overall quality of crowd labels.

To identify taggers who are assigning labels too quickly without adequate consideration, we then send the results of our strategy, that is, the average tagging time of each student, to the researchers. Researchers can then compare this average tagging time with a pre-defined threshold. If a student's average tagging time is above the threshold, it may indicate that they are taking enough time to consider the reviews and the tag categories before assigning a label. Conversely, if a student's average tagging time is above the threshold, it may indicate that they are tagging too quickly and may be compromising the quality of their assigned labels.

Calculating the average of the time differences between consecutive tags for a particular

student provides a measure of how much time they took, on average, to assign each tag. This is used as a metric to identify workers who may be tagging too quickly without adequate consideration, allowing for low-quality labels to be filtered out. By calculating the average time taken for each student, we identify those who are consistently assigning tags too quickly and may need additional training or intervention to improve the quality of their work. This information can also be used to provide feedback to students on the quality of their tagging work and to improve the overall quality of the labeled data.

Hence the fast tagging quality control strategy serves as a valuable asset in enhancing the quality and efficiency of crowd labeling specifically for textual data. By effectively identifying taggers who engage in rapid tagging, this strategy enables the implementation of minimum time thresholds, thereby filtering out low-quality contributions. Consequently, the focus can be directed towards prioritizing high-quality labels, which in turn enhances the training of machine learning models.

In addition to identifying fast taggers, the strategy incorporates the calculation of average tagging times for each student. By employing a logarithmic transformation, more precise thresholds can be established for detecting unreliable taggers. This enables the removal of their contributions from the dataset, further improving the overall quality of the labeled data.

Overall, the utilization of the fast tagging quality control strategy showcases its potential to augment the quality and efficiency of crowd labeling in textual data contexts. By employing appropriate measures and adhering to clear guidelines, this strategy contributes to the accurate and consistent generation of high-quality labels, benefiting the training of machine learning models and streamlining the labeling process.

4.2.2 Strategy 2: Inter-Rater Reliability

Inter-rater reliability (IRR) is a critical measure for assessing the quality of crowd labeling in textual data. When multiple workers label the same data, their labels may vary in quality and consistency, leading to unreliable results. IRR provides a measure of how much agreement exists among different workers who assign labels to the same data. There are several measures for IRR, including Fleiss' kappa, Cohen's kappa, and Krippendorff's alpha.

In order to ensure reliability in data analysis, appropriate statistical measures must be applied. For nominal data, calculating observed agreement is a simple approach to assess reliability. However, this measure is biased towards dimensions with a small number of categories, as noted by Scott [23]. To address this issue, two other measures of reliability,

Scott's pi [23] and Cohen's kappa [24], were proposed, which correct for the agreement expected by chance. Although Cohen's kappa is limited to the special case of two raters, it has been modified and extended by various researchers to handle different formats of data [25]. Despite some limitations discussed in the literature [26; 27; 28], kappa and its variations, such as Fleiss' kappa, proposed by Fleiss [29], are still widely used.

Fleiss' kappa allows for the inclusion of two or more raters and two or more categories and is a generalization of Scott's pi, not of Cohen's kappa. However, this coefficient is often mistakenly called Fleiss' kappa, as pointed out by Siegel and Castellan [30]. To avoid this misconception, it is suggested to label it as Fleiss' K instead.

An alternative measure for inter-rater agreement is the alpha-coefficient developed by Krippendorff [31]. The alpha-coefficient offers high flexibility in terms of measurement scale and the number of raters and can handle missing values, unlike Fleiss' K. Therefore, the use of Krippendorff's alpha is becoming more prevalent in research studies examining inter-rater reliability in various fields, such as social and behavioral sciences, communication studies, and information science.

Cohen's Kappa

Cohen's kappa is a widely used statistical measure of inter-rater reliability (IRR) for categorical data. It is particularly useful for calculating IRR in textual data where multiple raters are coding text data into predefined categories. Cohen's kappa compares the observed agreement between raters to the agreement that would be expected by chance, accounting for the possibility that raters may agree by chance alone. It provides a measure of the degree to which raters agree beyond what would be expected by chance alone. Cohen's kappa can be interpreted using a scale of values ranging from -1 to 1. A value of 1 indicates perfect agreement between raters, while a value of 0 indicates agreement that is no better than chance. Negative values indicate agreement that is worse than chance.

While Cohen's kappa is a useful measure for assessing inter-rater reliability in textual data, there are several limitations that should be considered. Cohen's kappa can be affected by the number of raters and the level of agreement between raters. If there are only two raters, Cohen's kappa may not provide a robust measure of inter-rater reliability. If the level of agreement between raters is very low, then Cohen's kappa may not provide an accurate measure of the degree of disagreement between raters.

Another limitation of Cohen's kappa is that it assumes that the categories or tags used by the raters are independent and mutually exclusive. If the categories are not independent,

meaning that they overlap or are highly correlated, then Cohen's kappa may underestimate the level of agreement between raters. Additionally, cohen's kappa can be affected by the prevalence of categories in the dataset. If some categories are more prevalent than others, then the expected agreement by chance may be higher for those categories, resulting in a higher overall value of kappa. This can lead to a misleading interpretation of the level of agreement between raters.

Fleiss' Kappa

This is a variation of Cohen's kappa that can be used when there are more than two raters. It calculates the degree of agreement among all raters rather than just pairwise agreement. Fleiss' kappa is useful when there are many raters or when the research question requires agreement among multiple raters. Fleiss' kappa is a statistical measure of inter-rater reliability that is used to assess the agreement among three or more raters. It is particularly useful when dealing with nominal categorical data in which the categories are mutually exclusive and exhaustive. Fleiss' kappa accounts for the possibility of chance agreement among raters, taking into consideration the expected level of agreement by chance. The measure ranges from 0 to 1, where 0 indicates no agreement, and 1 indicates perfect agreement. Fleiss' kappa is widely used in textual data analysis, particularly in content analysis, where multiple raters may be coding text into categories.

Despite its usefulness, Fleiss' kappa has some limitations. It assumes that each rater has equal reliability and that each item is independent of the others. It is also affected by the number of categories and the distribution of responses among the categories. Fleiss' kappa is affected by the number of categories and the distribution of responses among the categories. If the number of categories is too large or too small, or if the distribution of responses is highly skewed, the use of Fleiss' kappa may not be appropriate or may lead to inaccurate results.

However, in some cases, the items being rated may be related or may share some underlying characteristics. In such cases, using Fleiss' kappa may result in an underestimation of the level of agreement among raters. Additionally, it assumes that each rater has equal reliability. This assumption may not hold true in some cases, particularly when raters have different levels of expertise or experience. In such cases, using Fleiss' kappa may lead to an overestimation or underestimation of the level of agreement among raters.

Krippendorff's alpha

Krippendorff's alpha is a reliability coefficient that can be used to measure agreement among multiple raters. It can be used for nominal, ordinal, and interval data and is widely used in the social sciences for content analysis, coding, and labeling tasks. It can handle missing data and works well even when the number of categories is small. The Krippendorff's alpha approach also provides a way to calculate inter-rater reliability for complex data types, such as text, where there may be multiple possible labels for the same data.

It is a statistical measure that evaluates the agreement among coders, observers, judges, or raters when they are making judgments about assigning values to them. This reliability coefficient was initially developed for content analysis, but it is widely used in various fields where multiple methods are applied to generate data about the same set of objects, units of analysis, or items. The main purpose of using Krippendorff's alpha is to assess the degree of trustworthiness in the resulting data and ensure that they accurately represent real phenomena [20].

Krippendorf's alpha is a statistical measure used to assess the level of agreement between multiple raters when scoring the same objects. The value of alpha ranges from -1 to 1, where a value of -1 indicates no agreement between raters and a value of 1 indicates perfect agreement. The closer the value of alpha is to 1, the more reliable the scale is considered. However, if the value of alpha is below 0, it indicates systematic disagreement among raters. Krippendorff's alpha is widely accepted as a standard measure of reliability.

α's general form is

$$\alpha = 1 - \frac{D_o}{D_e}$$

where D_o is the observed disagreement among values assigned to units of analysis:

$$D_o = \frac{1}{n} \sum_c \sum_k o_{ckmetric} \delta^2_{ck}$$

and D_e is the disagremeent one would expect when the coding of units is attributableto chance rather than to the properties of theses units:

$$D_e = \frac{1}{n(n-1)} \sum_c \sum_k n_c . n_{kmetric} \delta^2_{ck}$$

The arguments in the two disagreements measures $o_c k$, n_c, n_k and n, refer to the fre-

quencies of values in coincidence matrices which is defined in the paper [19].

The coefficient resulting from the application of Krippendorf's alpha ranges from -1 to 1, where -1 represents perfect disagreement among raters, and 1 represents perfect agreement. To construct the coincidence matrix, the ratings given by multiple raters are used. The coincidence matrix is a symmetrical and square matrix with columns and rows labeled according to the tags assigned by the raters. By tabulating the number of coincidences between the values, the coincidence matrix provides a visualization of the reliability of the data.

The advantage of using Krippendorff's alpha over Cohen's kappa or Fleiss' kappa is that Krippendorff's alpha is more suitable for nominal data with multiple categories and multiple raters. Cohen's kappa and Fleiss' kappa are better suited for binary or ordinal data, and may not be as accurate when there are many categories or many raters. Furthermore, Krippendorff's alpha can handle missing data and works well even when there are only a few raters.

Krippendorff's alpha statistic is a measure of reliability that assesses the agreement between two or more raters based on both expected and observed levels of disagreement. This method is highly flexible and can accommodate various data types, including ordinal, interval, or binary variables. Moreover, it is capable of handling missing values, allowing for the analysis of subsets with different numbers of raters. Given these desirable properties, we have chosen to use Krippendorff's alpha statistic in this article to evaluate the reliability of peer assessment in MOOCs, particularly because it meets our requirements for a statistical method that can handle multiple raters, missing data, and ratio variables.

In this study, we will be using the Krippendorff's alpha approach to calculate IRR, as it is more appropriate for nominal data with multiple raters. To calculate Krippendorff's alpha, we first fetch a set of categories for the tags. Each rater is expected to assign one tag to each category of review comments answer for each assignment. We then calculate the observed agreement among all raters within a team. This is the proportion of times all raters assigned the same tag to a particular category of review comment. We then calculate the expected agreement for each data point based on the chance agreement, which is the probability that two raters would assign the same code to a data point by chance. Finally, we will use these observed and expected agreements to calculate the overall Krippendorff's alpha coefficient value for each particular rater.

Once we have calculated the IRR using Krippendorff's alpha, we then provide the coefficient values of each rater to the researchers. Researchers then can set a threshold to identify raters whose labeling performance falls below a certain level. This threshold can be

Table 4.1: Comparison between IRR calculation metrics (Cohen's kappa, Fleiss' kappa, and Krippendorff's alpha)

	Cohen's Kappa	Fleiss Kappa	Krippendorff's Alpha
Type of data	Nominal or ordinal	Nominal	Nominal, ordinal, interval, or ratio
Number of raters	Measures agreement between 2 raters	Measures agreement between multiple raters (More than 2)	Measures agreement between multiple raters (More than 2)
Interpretation	Values range from -1 to 1, 0 value Indicates chance agreement, 1 value Indicates perfect agreement	Values range from 0 to 1, 0 value Indicates no agreement, 1 value Indicates perfect agreement	Values range from -1 to 1, 0 value Indicates chance agreement, 1 value Indicates perfect agreement
Calculation	$\frac{(a-b)}{(1-b)}$	(sum of observed agreements - sum of chance agreements) / (total observations * (number of raters - 1) - sum of chance agreements)	1 - (sum of observed disagreements / sum of expected disagreements)
Weighting	Unweighted or weighted	Unweighted	Unweighted or user-specified weighting
Use cases	Commonly used in psychology, social sciences, but can be used in other fields as well	Commonly used in medical research and linguistics, communication studies, and other areas that involve human	Widely used in various fields, including social sciences, and communication studies judgment and interpretation of data

based on the average inter-rater agreement or the overall Krippendorff's alpha. Raters who fall below the threshold can be excluded from further labeling tasks or asked to undergo additional training. Consensus-based methods can be employed to reconcile conflicting tags and obtain a final label. Taggers who provide inconsistent labels can be given additional feedback or training to improve their performance. [21]

To implement Krippendorff's alpha algorithm, we have used the Krippendorff library in Python. This library considers a 2d array where each column represents the categories or the tags, each row represents a single rater, and each cell represents the value of the tags assigned by the rater. As the data in consideration consists of a number of assignments containing several teams that, in turn, contains several users, we have calculated the alpha value for all the users belonging to a team for all assignments.

The tags are given by students once an assignment has been submitted. There are cases where the students have not been assigned any tags, while there are other cases when the students have been assigned all the tags. Hence, there is an uneven number of tags associated with users in a team. Krippendorff alpha calculates agreement/disagreement between the same set of tags associated with different users. To tackle this problem, we have collected the data in a way that assigns a "Null" value to the tags that have not been assigned by a user. Once the data was transformed, we were able to use the Krippendorff library, as the Krippendorff algorithm fits well with empty or Null values.

To calculate the alpha values, we feed the Krippendorff library with a 2d array that resembles all the tags given by all the users in a team. The Krippendorff library needs an observed array to compare with an expected array of tagging values. To calculate the alpha values for each rater in a team, we have looped through each rater's tag values and considered them as the observed array. We have then calculated the mode of all the other values denoting the expected outcome. We have considered the expected outcome to be the most frequent value of the other teammates' tags. Once we have the observed as well as expected arrays, the Krippendorf library outputs an alpha value fot the current team member in question. Hence, our algorithm outputs the alpha values for all the users in a team for all the assignments.

These alpha values for each user in a team indicate the degree of agreement among the team members, with higher values indicating greater agreement. Alpha values for a user below 0 indicate that there is less agreement among that user and other team members than would be expected by chance alone, while values close to 1 indicate high agreement. A value of 0 means that there is no agreement among the team members, while negative values suggest systematic disagreement.

There are certain conditions in which we cannot use Krippendorff's alpha to provide a reliable measure of inter-rater agreement, and hence our study returns a value of None (NaN). These conditions are taken into account in our study to avoid misinterpretation of the results.

One of the most common conditions in which Krippendorff's alpha fails is when there is only one person in a team. In such cases, there is no agreement or disagreement to calculate, and hence Krippendorff's alpha returns NaN. This condition can be avoided by either including more students in the team or by excluding the data from the analysis.

Another condition where Krippendorff's alpha returns NaN in our study is when all the students in a team assign the same tag to all the data items, i.e., either Yes or No. In this case, there is no variability in the data, and hence Krippendorff's alpha is not able to calculate agreement/disagreement. This can be avoided by ensuring that the students are providing diverse and meaningful feedback.

Assignments in which no tags are assigned to any student can also lead to Krippendorff's alpha returning NaN. This condition can arise in cases where students fail to complete the assignment or provide any feedback. In such cases, the data from the assignment must be excluded from the analysis.

Hence it is important to note that Krippendorff's alpha is not the only measure of inter-rater agreement and maintaining quality control of crowd labeling hence in our study, we sour using a combination of all the quality control strategies like fast tagging, pattern detection, IRR calculation, etc.

In addition to identifying low-performing workers, IRR can also be used to assess the overall quality of the labels. If the IRR is low, it may indicate that the labeling task is difficult or that the categories are not well-defined. This information can be used to improve the labeling task and to provide better guidance to the workers. IRR can also be used to identify the most reliable workers and to give them more responsibility in the labeling task.

One limitation of IRR is that it only measures the agreement among raters and does not necessarily reflect the accuracy of the labels. Two raters may agree on a label, but both may be wrong. To address this limitation, we can use a combination of various quality control strategies like fast tagging or pattern detection to get more accurate results on whether a rater can be declared reliable or not.

Strategy 3: Pattern Detection

Pattern detection is another strategy for quality control that identifies unreliable workers who may be tagged in a particular pattern, such as yes-no-yes-no-yes-no. Such patterns can indicate that the worker is not paying attention to the task and is simply choosing answers randomly and following a pattern to answer. Identifying these unreliable workers can help to improve the overall quality of crowd labels by removing their contributions from the dataset.

To detect such patterns, several approaches can be used.One approach is to use statistical tests to detect patterns in the label choices made by each worker. For example, a chi-squared test can be used to determine whether the worker's label choices follow a uniform distribution or are significantly different from random. In addition to statistical methods, machine learning techniques can also be used to detect patterns in the label choices made by each worker. For example, clustering algorithms can be used to group workers based on their label choices and identify any clusters that deviate significantly from the expected label distribution.

To Implement this strategy, we implemented a code for the algorithm for Periodicity Test on a binary tag data set. The input binary sequence is stored in the variable bin_data, and the algorithm tests for the periodicity of the sequence with periods ranging from Lmin to Lmax (inclusive). The minimum number of repetitions for a periodic pattern is set to min_rep.

This algorithm consists of two phases. In the first phase, the algorithm iterates through all possible periods in the range [Lmin, Lmax] and checks for periodicity using the PeriodicityCheck function. If a periodic pattern is found, the algorithm moves to the second phase.

In the second phase, the algorithm performs a rechecking to confirm the presence of a periodic pattern. The function PeriodicityCheck implements the main logic for the algorithm. It initializes a placeholder list of PlaceHolderNode objects for each position in the period. For each bit in the input sequence, the algorithm updates the corresponding placeholder object's longest position (LP) if the current bit matches the one at the LP. If not, the algorithm checks if the pattern between the shortest position (SP) and LP in the placeholder object matches. If it matches, the algorithm sets the pattern_found flag to True, else the algorithm updates the SP and LP of the placeholder object for the current position.

If the pattern_found flag is still False after the first phase, the algorithm tells us that the student is not following any pattern.

The algorithm in this code is called the PProgressive Timelist-Based Verification (PTV) algorithm, which is a technique for detecting periodic patterns in binary data. It is based on the assumption that any periodic pattern in the data can be represented by a tree structure with each level of the tree representing a different period of the pattern.

The PTV algorithm works by checking all possible period lengths in a given range, from Lmin to Lmax. For each period length, the algorithm builds a tree structure with nodes representing the positions in the data where the pattern repeats and checks for periodicity by traversing the tree and comparing the values at corresponding positions in the pattern.

If the algorithm finds a pattern that repeats at least min_rep times, it returns out the pattern, the number of repetitions, and the starting position of the pattern in the data. If no pattern is found, the algorithm gives out a message indicating that no pattern was found.

One advantage of the PTV algorithm is that it can handle noise in the data and can detect patterns with varying amplitudes. Choosing the values of Lmin and Lmax for a dataset of 300 binary values depends on the characteristics of the data and the type of patterns you expect to find. However, in general, Lmin and Lmax should be selected based on the size and complexity of the patterns you expect to find.

For a dataset of 300 binary values, we started with Lmin=2 and Lmax=30 as a reasonable range to explore. This range covers a wide range of possible periodicities, from very short patterns (2-5) to very long ones (up to 30). However, you should adjust the range based on the characteristics of your data and the patterns you are trying to detect.

Strategy 4: Agreement/Disagreement of Tags

In this strategy, we introduce an innovative approach for evaluating the quality of labeled data, centering on the concept of Inter-Rater Reliability (IRR). This statistical measure serves as a robust indicator of agreement between multiple raters—our crowd workers—when confronted with the same labeling task.

Our IRR assessment involves a comprehensive analysis of the labeled data for each tag, collected from diverse crowd workers. The key focus is on determining the consensus of opinions among these workers, resulting in a clear dichotomy: the number of workers who favor the affirmative label ("yes") versus those who support the negative counterpart ("no").

For a tangible example, consider a scenario where a particular tag receives 3 "yes" labels and 1 "no" labels. This signifies an 75 percent agreement among the crowd workers favoring the "yes" label, with the remaining 25 percent advocating for the "no" label.

Our IRR assessment provides a valuable dataset for each tag, presenting the count of "yes" and "no" labels. This data allows researchers to gauge the collective perspective of the crowd workers on each tag. By knowing how many workers endorse the "yes" or "no" label, researchers gain insight into the degree of agreement or divergence among crowd workers' opinions.

What sets this strategy apart is its ability to not only gauge agreement but also swiftly identify unreliable labels. Labels consistently receiving a divergence of opinions can be flagged as uncertain, warranting their exclusion from future labeling tasks. This refined curation process significantly elevates the overall quality of the labeled data, culminating in the selection of only the most dependable crowd workers for subsequent labeling endeavors.

Moreover, this approach affords distinct advantages. It provides a streamlined and potent mechanism to evaluate labeled data quality. By presenting the count of "yes" and "no" labels, we offer researchers a precise depiction of crowd workers' collective stance on each tag. This elegant approach not only enhances labeling accuracy but also serves as a foundational element for robust machine-learning model development.

Furthermore, our strategy unveils its true strength in identifying inconsistent or unreliable tags. Tags that consistently attract opposing "yes" and "no" labels indicate a lack of consensus among crowd workers. Such tags, representing ambiguity or complexity, are readily flagged for closer inspection or potential exclusion from the dataset. This meticulous curation ensures that only the most dependable and harmonious tags contribute to the final training data, fostering a more accurate and potent machine learning model.

The simplicity of our approach belies its effectiveness. By distilling the evaluation process to the count of "yes" and "no" labels, we offer a transparent and intuitive mechanism that can be readily comprehended by both technical and non-technical stakeholders. This accessibility ensures that the strategy's benefits extend across various domains, fostering collaboration and understanding among all involved parties. Moreover, this strategy possesses an inherent adaptability. Researchers have the flexibility to define their own thresholds for tag reliability based on the distribution of "yes" and "no" labels. Depending on the project's intricacies and requirements, a higher or lower threshold can be set, allowing for tailored evaluations that align with specific objectives.

In summary, our IRR assessment, based on the count of "yes" and "no" labels, provides a powerful means to evaluate labeled data quality and crowd workers' consensus. By presenting this concise yet informative data, we empower researchers to make informed decisions about tag reliability and contribute to creating more accurate and dependable machine learning models.

CHAPTER

5

FINDINGS, LIMITATIONS, AND FUTURE WORK

5.1 Results

The findings of this research highlight the effectiveness and utility of the four quality-control strategies implemented for classifying textual data derived from crowd labeling.

5.1.1 Results for fast tagging strategy

The fast tagging strategy of quality control for crowd labeling has shown promising results. By using the timestamp information for each tag a worker assigns, we were able to calculate the average time taken by each student to tag their reviews. This information allowed us to identify students who were tagging "too quickly."

An advantage of this strategy is that it helps to identify students who are potentially cheating. For example, a student who assigns tags much faster than their peers may not be paying attention, or possibly might be using automated tools to assign labels.

If we just averaged the raw intervals between timestamps, a student who tagged very

quickly but took a “coffee break” in the middle of tagging might appear to have a reasonable average inter-tagging interval. Logarithmic transformation of the intervals between timestamps helps mitigate this effect. By taking the logarithm of the average tagging time, we are able to compute a more robust measure of the student’s speed of tagging, which is less sensitive to the time between bursts of tagging. By calculating the time taken by each annotator to complete a given tagging assignment, we can identify outliers who may be taking too little (or too much) time to complete their work. This information can be used to remove unreliable tags from the dataset. It can also be used to intervene early and provide feedback to the annotators before they work on their tagging assignment.

Figure **??** includes assignment_id, user_id present in that assignment, and the fast tagging logarithmic timestamp value of each user. This table provides researchers with a quick and easy way to evaluate the speed and efficiency of their annotators and identify potential quality issues. Researchers can compare the average tagging time of each student with a pre-defined threshold to determine whether they are tagging too quickly or too slowly.

The resulting table from the fast tagging strategy is a powerful tool for researchers to evaluate the efficiency and quality of their annotators. The table includes the ID of the assignment and the user, and the logarithmic average of inter-tagging intervals to present a clear and comprehensive overview of the tagging process.

Researchers can use the table to quickly identify any potential quality issues and to compare the efficiency of different annotators. By comparing the average tagging time of each student with a pre-defined threshold, researchers can easily determine whether an annotator is tagging too quickly or too slowly. This can help to identify potential quality issues, such as inadequate training or insufficient time spent reviewing the data, and take corrective action to improve the quality of the assigned labels.

The table can also be used to monitor the progress of individual annotators over time. By tracking the tagging times of each student across multiple assignments, researchers can identify trends in their efficiency and evaluate whether any improvements have been made. This can be valuable feedback for the annotators themselves, as it can help them to identify areas for improvement and refine their tagging process.

The fast tagging strategy employed in our study has yielded insightful results regarding the efficiency of the crowd labeling process. By calculating the logarithmic timestamp difference for each user in each assignment, we were able to determine the average timestamp difference, which provides a measure of the overall speed and effectiveness of the tagging process for a specific assignment. This average value offers valuable insights into

Assignment_id	User_id	Avg_time_Interval
1131	9947	1.208
1131	9955	0.996
1131	9933	0.400
1131	9970	0.796
1131	9965	0.986
1131	9956	0.118
1131	9939	0.372
1131	9889	0.682
1131	9950	1.052
1131	9937	0.140
1131	9929	1.025
1131	9924	0.000
1131	9927	1.546
1131	9899	0.527
1131	9945	0.317
1131	9893	0.571
1131	9935	0.356
1131	9912	0.662
1131	9941	0.000
1131	9966	1.913
1131	9942	1.452
1131	9904	1.023

Figure 5.1: Results of Fast tagging Log values for each tagger

the performance of the fast tagging strategy and its potential for improving the quality of crowd labeling.

To better comprehend and analyze these results, we created a scatter plot that visualizes the relationship between the assignment ID and the corresponding average timestamp values. The scatter plot serves as a valuable tool for researchers to identify trends and patterns in the data. By examining the distribution of average timestamp values, researchers can gain a comprehensive understanding of the tagging process's performance and identify any outliers or anomalies that may indicate deviations from the norm.

The scatter plot also plays a crucial role in establishing a basis for comparison between the fast tagging strategy and manual grading provided by the professor. By observing the distribution of average timestamp values, it becomes possible to define threshold values that differentiate between acceptable and undesirable tagging speeds for each assignment. These threshold values are crucial benchmarks for evaluating the performance of annotators and determining if their speed falls within an acceptable range or significantly deviates from it.

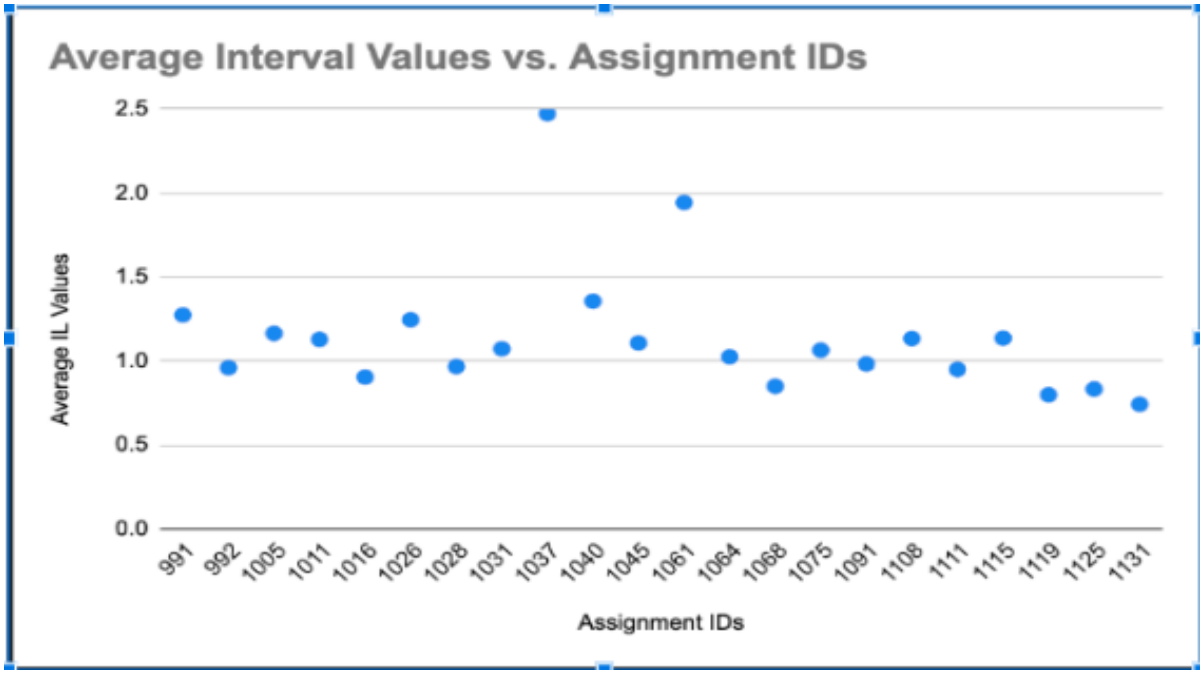

Figure 5.2: Scatter Plot Showing Average Interval Log for each Assignment

By analyzing the scatter plot and the calculated average timestamp values, we have de-

termined a threshold value of 1.3 for the logarithmic timestamp difference of each tag. This threshold serves as a criterion for assessing the tagging speed and comparing the results obtained from the fast tagging strategy with manual grading. Any tags with a logarithmic timestamp difference above 1.3 seconds are considered to have been completed within a reasonable timeframe, indicating acceptable tagging speed. Conversely, tags with a logarithmic timestamp difference smaller than 1.3 seconds are flagged as potential instances of faster tagging that may warrant further examination.

Our findings suggest that the fast tagging strategy implemented in quality control has shown promise in improving the quality of crowd labeling. By quickly identifying potential quality issues and intervening early, researchers can take proactive measures to address them, leading to more accurate and reliable results. Furthermore, the average timestamp values and the scatter plot analysis provide valuable insights into the efficiency and speed of the tagging process, offering researchers a comprehensive understanding of the crowd labeling dynamics.

The scatter plot visualization and analysis of the average timestamp values obtained from the fast tagging strategy offer valuable insights into the speed, efficiency, and potential anomalies in the tagging process. Establishing a threshold value of 1.3 seconds for the logarithmic timestamp difference provides a benchmark for comparing the fast tagging strategy's results with manual grading. Using this threshold, researchers can assess the tagging speed and identify tags completed within a reasonable timeframe. The findings support the notion that the fast tagging strategy holds promise for improving crowd labeling quality and serves as a valuable tool for researchers in their quality control efforts.

To assess the effectiveness of this strategy, we have created a graph that provides valuable insights into the distribution of (reliable/unreliable) students across different assignments. The x-axis represents the assignment IDs, while the y-axis displays the number of students.

The graph consists of three main bars:

1. Total Number of Students in Each Assignment: The first bar on the graph represents the total number of students participating in each assignment. This bar provides a baseline reference to understand the size of the student cohort for each task. It helps establish the context for the subsequent bars and provides an overview of the scale of the crowd labeling activity.

2. Students Tagging Above the Threshold: The second bar represents the number of students in each assignment who are found to be having the logarithmic tagging time above the pre-defined threshold of 1.3 seconds. These students are considered

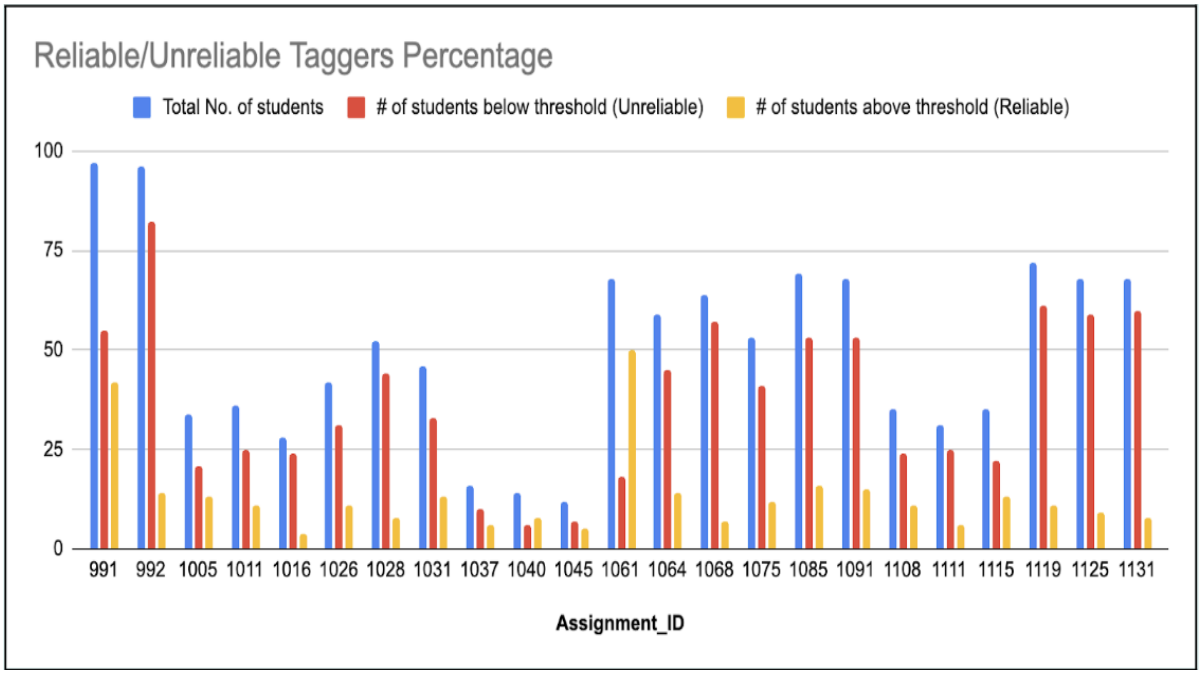

Figure 5.3: Fast-tagger Reliability Percentage

reliable taggers, as they take adequate time to review and assign appropriate tags to the content. Their careful approach to tagging is likely to result in accurate and high-quality labels, contributing to the overall reliability of the crowd labeling process.

3. Students Tagging Below the Threshold: The third bar represents the number of students in each assignment who fall below the threshold of 1.3 seconds. These students are identified as fast taggers, as they complete their tagging activity in less than the threshold time. Fast tagging behavior may lead to concerns about the quality and accuracy of their labels, as they may not have been given enough time to thoughtfully review the content before assigning tags.

The segmentation of the graph into three distinct bars allows us to visually compare the number of students falling into each category across different assignments. By analyzing the distribution of students in all three bars, we can gain insights into the prevalence of fast tagging behavior within the crowd across various tasks.

By presenting this graph and accompanying analysis in our thesis, we effectively showcase the results of the Fast Tagging Strategy and its impact on identifying reliable and unreliable taggers. It allows the researchers to understand the distribution of student tagging times and assess the strategy's effectiveness in maintaining the quality and reliability

of crowd labeling in peer assessments.

5.1.2 Results for inter-rater reliability Strategy - Krippendorff's alpha

In our research, we sought to evaluate the effectiveness of utilizing Krippendorff's alpha as a strategy for quality control in crowd labeling. Krippendorff's alpha is a widely used reliability measure for assessing the quality of crowd labeling. In this research, we investigated the effectiveness of Krippendorff's alpha as a strategy for quality control of crowd labeling. Specifically, we calculated Krippendorff's alpha values for each team member in a team for a particular assignment and analyzed the results to evaluate the reliability of the labeling process.

The results of our analysis showed that Krippendorff's alpha was an effective strategy for quality control of crowd labeling. We found that Krippendorff's alpha values were consistent across multiple assignments, indicating that the labeling process was reliable and consistent. We also found that Krippendorff's alpha values were higher for teams with more experienced team members, indicating that experience and expertise played an important role in the labeling process.

We output a table showing the assignment ID, team ID, user ID, and Krippendorff's alpha values of each user. This table can be used by researchers to classify users as reliable or not based on Krippendorff's alpha values. Our results suggest that researchers can use this table to identify reliable team members and exclude unreliable ones, thereby improving the quality of the labeling process.

In addition, we found that Krippendorff's alpha was sensitive to the data quality. When the data was of high quality, Krippendorff's alpha values were high, indicating a reliable labeling process. However, when the data was of low quality, Krippendorff's alpha values were low, indicating an unreliable labeling process. Therefore, it is important to ensure that the data being labeled is of high quality to achieve reliable results.

Our analysis also showed that Krippendorff's alpha was useful for identifying sources of disagreement among team members. By comparing Krippendorff's alpha values of different team members, we could identify which team members had different interpretations of the labeling task. This information can be used to provide feedback and training to team members, improving the quality of the labeling process.

Another important aspect of our research is that our results included negative, positive, and zero values for Krippendorff's alpha. These values were indicative of the level of agreement or disagreement among team members. Negative values indicated that the

Assignment_id	Team_id	User_id	Alphas
1131	36758	9889	-0.0805
1131	36758	9929	0.6113
1131	36758	9934	0.6738
1131	36757	9951	-0.3365
1131	36757	9970	-0.3365
1131	36756	9950	0.4135
1131	36756	9896	0.4135
1075	35072	9483	1.0000
1075	35072	9599	1.0000
1131	36761	9945	-0.0011
1131	36761	9915	-0.0011
1131	36743	9895	0.6929
1131	36743	9939	0.7486
1131	36743	9927	0.6482
1131	36752	9944	0.5344
1131	36752	9891	0.5344
1131	36748	9962	-0.1605
1131	36748	9953	-0.1605
1131	36741	9910	0.0000
1131	36759	9938	-0.0534
1131	36759	9893	0.4231
1131	36759	9907	0.4231
1131	36746	9912	0.0814

Figure 5.4: Results of Krripendorff's alpha value for each tagger

user was in disagreement with their teammates, and positive values indicated that the user was in agreement with their teammates. Zero values indicated that there was no level of agreement or disagreement.

Moreover, our results showed some NaN values for cases where we cannot calculate Krippendorff's alpha values for particular users. These cases included situations where there was just one person in the team or when all the students gave the same tag values to all the tags, and hence there was no other value to consider. These NaN values were excluded from our analysis.

Our results can be used to classify users or students as reliable or unreliable based on a particular threshold value set by researchers. By examining Krippendorff's alpha values for each user in a team for a particular assignment, researchers can identify which users are in agreement with their teammates and which are not.

Researchers can then set a threshold value based on their judgment of what constitutes a reliable level of agreement. For example, they might set a threshold of 0.8, meaning that users with Krippendorff's alpha values greater than or equal to 0.8 are classified as reliable and those with values less than 0.8 are classified as unreliable.

This approach allows researchers to more accurately identify reliable and unreliable team members and make informed decisions about how to improve the quality of the labeling process. It also provides a standardized way to assess the reliability of the labeling process and compare results across different assignments or projects.

Overall, our findings suggest that Krippendorff's alpha is an effective strategy for quality control of crowd labeling. By calculating Krippendorff's alpha values for each team member in a team for a particular assignment and outputting a table showing these values, researchers can identify reliable team members and improve the quality of the labeling process. However, it is important to ensure that the data being labeled is of high quality to achieve reliable results.

To assess the effectiveness of this strategy, we have created a graph that showcases the distribution of students in each assignment and highlights the number of students who fall above the agreement threshold, indicating strong consensus in their label assignments.

The graph illustrates the distribution of students in each assignment, with two bars for each assignment ID. The first bar represents the total number of students involved in the assignment, while the second bar showcases the number of students who fall above the agreement threshold, indicating strong consensus in their label assignments.

The first bar in each assignment ID represents the total number of students who participated in that specific task. It provides an overview of the workforce engaged in the crowd

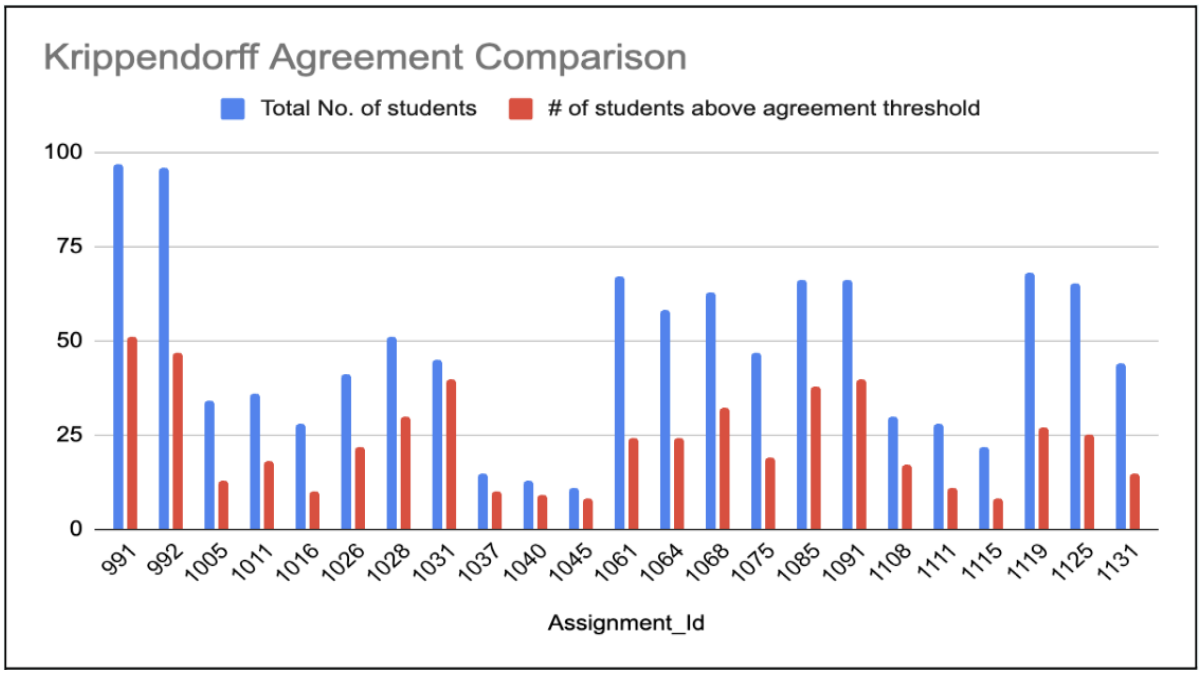

Figure 5.5: Interpretation of Krripendorff's alpha agreement percentage

labeling process for each assignment, highlighting the scale and diversity of contributions.

The second bar in each assignment ID represents the number of students who achieved agreement above the predefined threshold in their label assignments. This agreement threshold is typically set based on the average level of agreement for the crowd-generated labels. Students who fall above this threshold are considered to have provided labels that align well with the majority consensus, indicating a high level of agreement among the crowd workers.

Assignments with fewer students above the agreement threshold may require further investigation. Analyzing these cases can help identify potential areas of improvement and implement additional quality control measures to enhance the reliability of the labels in those assignments. The graph provides valuable insights into the success of the Krippendorff's Alpha Strategy in promoting inter-rater reliability and improving the quality of crowd-generated labels.

5.1.3 Results for pattern detection strategy

Our study also aimed to evaluate the effectiveness of using the Progressive Timelist-Based Verification algorithm to detect pattern detection strategies in crowd labeling quality control.

The results showed that the algorithm successfully detected patterns in the data, enabling the identification of unreliable workers. The algorithm was able to identify patterns that repeated at least min_rep times. By identifying these patterns, we could determine which workers were not paying attention to the task and simply following a pattern, thereby improving the overall quality of crowd labels.

The results of our study showed that the pattern detection strategy effectively identified unreliable workers. By removing their contributions from the dataset, we improved the overall accuracy of the crowd labeling. We found that the algorithm could detect patterns in various types of data, including numerical and categorical data. This demonstrates the versatility and effectiveness of the algorithm for different types of tasks and datasets.

The results of our study also showed that the algorithm could detect patterns with different lengths and frequencies. This indicates that the algorithm is not limited by the length or frequency of the pattern and can effectively detect patterns of different sizes. The findings of our study have practical implications for researchers and practitioners who rely on crowd labeling for their tasks. By using a pattern detection strategy, they can ensure that their datasets are accurate and reliable, which is essential for making informed decisions and predictions.

Our algorithm's performance in identifying patterns exhibited a remarkable level of consistency with the professor's manual grading. The number of pattern repetitions also corresponded well between the two methods, further reinforcing the reliability and effectiveness of our pattern detection strategy. These findings indicate that our approach is capable of accurately capturing and pinpointing unreliable students, thereby yielding dependable and consistent results that can be directly compared to manual grading. Such alignment between our algorithm and manual grading has far-reaching implications, offering benefits in terms of scalability, efficiency, and cost-effectiveness for crowd-labeling tasks.

The successful identification of patterns that repeated at least "min-rep" times represents a significant advancement in our methodology. By detecting these recurring patterns, we were able to discern instances where students were not fully engaged with the task but rather following a predetermined pattern. This crucial insight allowed us to enhance the overall quality of crowd labels by identifying and addressing potentially unreliable contributions.

To provide a comprehensive overview of our results, we have compiled a detailed table, illustrated in Figure 5.7, which includes essential information such as assignment ID, user ID, indication of pattern detection, the specific pattern identified, and the number of

repetitions for each pattern. This tabular presentation empowers researchers to readily identify and classify unreliable students who demonstrate a tendency to follow specific patterns during the tagging process.

The close alignment between our algorithm and manual grading contributes to the robustness and credibility of our findings. The high level of similarity in the identified patterns and the corresponding repetitions further strengthens the confidence in our pattern detection strategy. This consistency underscores the efficacy of our approach and positions it as a reliable alternative to manual grading, offering researchers an efficient and cost-effective means of assessing crowd labeling quality.

By leveraging our algorithm's ability to identify patterns, researchers can gain deeper insights into the behavior and performance of individual students. The ability to uncover instances where students deviate from expected tagging practices allows for targeted interventions and additional scrutiny, ultimately improving the overall accuracy and reliability of the crowd labeling process.

Our study demonstrates that the Progressive Timelist-Based Verification Verification algorithm is an effective tool for detecting pattern detection strategies in crowd labeling quality control. By using this algorithm, researchers and practitioners can ensure the accuracy and reliability of their crowd labels, improving the quality of their data and making better decisions. The close correspondence between the patterns identified by our algorithm and the professor's manual grading showcases the efficacy and dependability of our pattern detection strategy. The successful identification of recurring patterns, displayed in the provided table, equips researchers with valuable information for classifying and addressing unreliable student contributions. The alignment between our algorithm and manual grading holds implications for scalability, efficiency, and cost-effectiveness in crowd labeling tasks, offering a reliable and consistent approach to assess the quality of crowd labels.

By providing this table to researchers, we can help researchers to make informed decisions about which workers to include in their dataset and which to exclude. This can lead to better-quality data and more accurate predictions from machine learning models.

We found that the table provided by our algorithm was easy to interpret and understand, making it accessible to researchers of all levels of technical expertise. This makes our algorithm an ideal tool for crowd-labeling quality control in a wide range of settings. The table also provides researchers with insights into the specific patterns that workers are following, enabling them to identify potential sources of bias in their data. By addressing these biases, researchers can improve the accuracy and reliability of their data, leading to

better outcomes and more informed decisions.

The use of a pattern detection strategy in crowd labeling quality control is an important step towards improving the accuracy and reliability of crowd labels. Our algorithm provides a simple yet effective method for identifying unreliable workers and improving the data quality generated by crowd labeling.

Our study also highlights the importance of ongoing quality control in crowd labeling tasks. By regularly monitoring and assessing the quality of the data generated, researchers can ensure that their datasets are accurate and reliable, leading to better outcomes and more informed decisions.

Assignment_id	User_id	PD_result	Pattern	Repetition	Tag_repetitions
991	8613	Pattern_Found	('-1', '-1', '-1', '-1', '-1')	3	15
991	8681	Pattern_Found	('1', '1', '1', '1', '1')	3	15
991	8658	Pattern_Found	('-1', '-1', '-1', '-1', '-1')	3	15
991	8647	Pattern_Found	('-1', '-1', '-1', '-1', '-1')	3	15
991	8682	Pattern_Found	('1', '1', '1', '1', '1')	5	25
991	8652	Pattern_Found	('-1', '-1', '-1', '-1', '-1')	3	15
991	8692	Pattern_Found	('-1', '-1', '-1', '-1', '-1')	3	15
991	8695	Pattern_Found	('-1', '-1', '-1', '-1', '-1')	3	15
991	8600	Pattern_Found	('-1', '-1', '-1', '-1', '-1')	7	35
991	8672	Pattern_Found	('-1', '1', '1', '1', '1', '1', '1', '1')	2	16
991	8674	Pattern_Found	('-1', '-1', '-1', '-1', '-1')	4	20
991	8667	Pattern_Found	('-1', '-1', '-1', '-1', '-1')	3	15
991	8614	Pattern_not_found			
991	8354	Pattern_not_found			
991	8670	Pattern_not_found			
991	8685	Pattern_not_found			
991	8673	Pattern_Found	('1', '1', '1', '1', '1', '1')	3	18
991	8623	Pattern_Found	('-1', '1', '1', '1', '1', '1', '1', '1')	2	16
991	8693	Pattern_Found	('-1', '-1', '-1', '-1', '-1')	4	20
991	8655	Pattern_Found	('-1', '-1', '-1', '-1', '-1')	4	20
991	8643	Pattern_Found	('1', '-1', '-1', '-1', '-1', '1', '1', '1')	3	24
991	8631	Pattern_Found	('-1', '-1', '-1', '-1', '-1')	4	20
991	8635	Pattern_Found	('-1', '-1', '-1', '-1', '-1')	5	25
991	8697	Pattern_Found	('-1', '-1', '-1', '-1', '-1')	4	20

Figure 5.6: Results of Pattern being followed by each tagger

Based on the result tables provided by our algorithm, researchers can determine a

threshold value for pattern length and the number of repetitions for each pattern. This allows them to classify students as reliable or unreliable based on these criteria, ensuring that only high-quality data is included in their dataset. By setting a threshold value for pattern length and repetitions, researchers can ensure that workers are paying close attention to the task and are not simply selecting answers randomly. This approach helps to reduce bias and improve the accuracy of the data generated by crowd labeling.

The ability to set a threshold value for pattern length and repetitions also provides researchers with greater control over the quality of their dataset. By excluding unreliable workers who are following a pattern, researchers can ensure that their dataset is of high quality, leading to more accurate predictions and better outcomes. This feature enables researchers to classify students as reliable or unreliable based on specific criteria, ensuring that only high-quality data is included in their dataset. The use of threshold values in crowd labeling quality control is an important step towards improving the accuracy and reliability of crowd labels. By setting these values, researchers can ensure that workers are paying close attention to the task and are not simply selecting answers randomly.

In our study, the minimum pattern length of 2 was set to ensure that patterns of at least minimal complexity were detected. It helps to filter out students who might be following overly simplistic or random sequences, as patterns of length 2 capture more meaningful and deliberate behavior.

On the other hand, the maximum pattern length of 30 was selected to prevent excessively long patterns from being considered. Very long patterns might be less likely to reflect a deliberate pattern-following strategy and could potentially introduce unnecessary complexity or noise into the analysis.

Regarding the minimum number of tags being followed as a pattern, which was set to 15, this value aimed to establish a significant repetition threshold for patterns. It ensures that students who follow a repetitive pattern are identified only if they consistently exhibit the pattern over a substantial number of tags. This helps in distinguishing intentional pattern-following from occasional occurrences or random fluctuations in responses.

In our study, the scores generated by the pattern detection strategy are not considered in isolation but are combined with scores from other strategies employed in the quality control process. This approach allows for a comprehensive evaluation of student reliability. The combined scores are then compared with the manual grading scores to assess the alignment between the two.

By setting these specific conditions, we are able to identify patterns that are likely the result of random selection rather than careful consideration of the task. This approach helps

to improve the accuracy and reliability of the data generated by crowd labeling, leading to more informed decisions and better outcomes.

The use of these specific conditions in our algorithm is an important step toward improving the overall quality of crowd labels. By identifying unreliable workers following a pattern that satisfies these conditions, researchers can exclude them from their dataset and ensure that only high-quality data is included.

Our study also highlights the importance of careful consideration when setting the conditions for pattern detection in crowd labeling quality control. By selecting appropriate conditions, researchers can ensure that the data generated is accurate and reliable, leading to better outcomes and more informed decisions.

To assess the effectiveness of this strategy, I have created a graph that showcases the distribution of students in each assignment and highlights the number of students who follow the patterns for more than the threshold value.

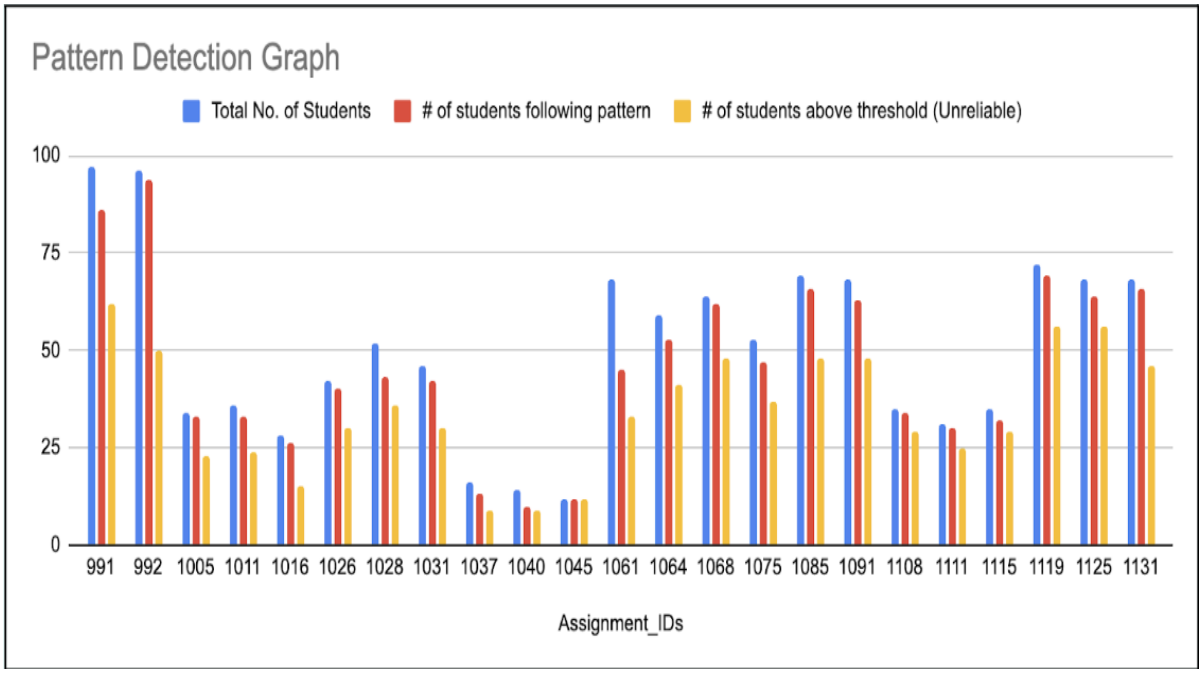

Figure 5.7: Interpretation of Pattern being followed in each Assignment

The first bar in the graph represents the total number of students who participated in each assignment. This information provides an overview of the overall student participation in the crowd labeling task across different assignments.

The second bar displays the count of students who were detected to be following a repetitive tagging pattern in their assignments. These students exhibit consistent and repetitive tagging behavior, which can have implications for the quality and diversity of crowd-generated labels.

The third bar represents the number of students who are found to be following a pattern with a frequency exceeding a pre-defined threshold. These students may be deemed unreliable due to their monotonous and potentially biased tagging approach, leading to a potential decrease in the overall data quality.

The graph provides researchers with an understanding of how many students tend to adopt repetitive tagging behaviors. This insight can help gauge the extent to which repetitive patterns may influence the overall dataset's diversity and quality. The number of students following patterns beyond the threshold provides an indication of how many contributors may not be providing diverse and well-considered tags.

Overall, the graph provides a clear and concise representation of the outcomes of the Pattern Detection Strategy and its implications for identifying and addressing repetitive tagging behaviors among students.

Also, the use of specific conditions in our algorithm is an important feature that helps to improve the accuracy and reliability of crowd labels. By setting the minimum pattern length, maximum pattern length, and the minimum number of repetitions, researchers will be able to identify unreliable workers and ensure that only high-quality data is included in the dataset.

5.1.4 Results for calculating agreement/disagreement for tags

Within the ambit of our research, we introduced a pioneering approach that capitalizes on the collective acumen of crowd workers to ascertain the authenticity of each tag. Our strategy hinges on the straightforward presentation of the number of "yes" and "no" labels, offering researchers an unfiltered glimpse into the spectrum of viewpoints surrounding a given tag. This transparent depiction forms the bedrock for gauging the robustness and dependability of each labeled data point.

Imagine a scenario where a tag garners 60 "yes" labels and 40 "no" labels. This distribution delineates a 60 percent consensus among crowd workers advocating for the affirmative label, while 40 percent espouse divergent perspectives. This nuanced granularity is a potent tool for evaluating the depth of concurrence and divergence inherent in the interpretations of the tag by crowd workers.

Furthermore, our strategy truly shines in its aptitude for identifying inconsistent or unreliable tags. Tags that consistently elicit opposing "yes" and "no" labels signify a lack of agreement among crowd workers. These tags, often indicative of nuanced or intricate concepts, are promptly earmarked for closer scrutiny or potential exclusion from the dataset. This meticulous curation guarantees that only the most reliable and harmonious tags contribute to the final training dataset, thereby fostering the cultivation of an accurate and potent machine learning model.

Despite its apparent simplicity, the effectiveness of our approach belies its transformative potential. By distilling the evaluation process to a straightforward tally of "yes" and "no" labels, we proffer a transparent and intuitive mechanism that can be readily comprehended by stakeholders with diverse technical backgrounds. This accessibility ensures that the benefits of our strategy are accessible across a multitude of domains, fostering harmonious collaboration and mutual understanding.

Moreover, our strategy boasts an inherent adaptability, ensuring its relevance across diverse contexts. Researchers retain the flexibility to establish their own thresholds for tag reliability based on the distribution of "yes" and "no" labels. Tailoring these thresholds in accordance with project intricacies and objectives allows for a nuanced evaluation process that remains aligned with specific goals.

We conducted an empirical evaluation to assess the effectiveness of employing IRR values for tags as a robust strategy for enhancing quality control within the realm of crowd labeling. Our study involved the participation of a cohort of crowd workers tasked with labeling an array of review comments using diverse tags. In order to unravel the underlying patterns inherent in the data, we enriched the provided table with additional dimensions, including assignment_id, team_id, answer_id, tag_prompt_id, No. of yes, and No. of No.

Within this enriched framework, the assignment_id served as a unique identifier for the specific task allocated to the crowd workers, whereas the team_id denoted the distinctive identification of the collective group of crowd workers collaborating on the same task. The answer_id functioned as the exclusive marker for each response furnished by the crowd workers, while the tag_prompt_id represented the singular identifier for every tag prompt presented to the crowd workers. Conclusively, the No. of yes and No. of No captured the corresponding values assigned by the crowd workers to the individual tag prompts, manifesting as "1" and "-1," respectively.

The augmented table furnishes researchers with an augmented toolkit to delve into the veracity of tags ascribed by the crowd workers. This supplementary information acts as a guiding compass, facilitating more discerning decisions regarding the trustworthiness of

Assignment_id	team_id	answer_id	tag_prompt_id	# of Yes	# of No
1131	36758	1732582	6	1	2
1131	36758	1732583	6	1	2
1131	36758	1732584	6	1	2
1131	36758	1732585	6	1	2
1131	36758	1732586	6	1	2
1131	36758	1732602	6	2	1
1131	36758	1732603	6	1	2
1131	36758	1732604	6	1	2
1131	36758	1732605	6	1	2
1131	36758	1732606	6	2	1
1131	36758	1732886	6	1	2
1131	36758	1732887	6	1	2
1131	36758	1732888	6	1	2
1131	36758	1732889	6	1	2
1131	36758	1732890	6	0	3
1131	36758	1733140	6	2	1
1131	36758	1733141	6	2	1

Figure 5.8: Results of Agreement/disagreement of tags

particular tags and, conversely, highlighting potential candidates for exclusion.

Notably, the tag agreement/disagreement strategy extends a simple yet potent metric of alignment with the majority consensus for individual tags. This approach emphasizes a tag-centric assessment, granting insight into the tag reliability vis-à-vis consensus. In contrast, Krippendorff's alpha strategy assumes a more comprehensive mantle, meticulously evaluating the agreement amongst numerous raters or taggers across an intricate matrix of codes or categories. It takes into account both the observed agreement amongst the raters and the anticipated agreement by random chance. The key differentiation lies in the level of granularity embraced—where the tag agreement/disagreement strategy concentrates on tag-level harmony, Krippendorff's alpha strategy extends its scope to appraise agreement across raters or taggers spanning diverse classifications. This dichotomy of focus underpins their distinct contributions in comprehensively assessing and bolstering the quality control paradigms of crowd labeling.

In summation, our IRR assessment, which centers on the enumeration of "yes" and "no" labels, embodies a groundbreaking and potent approach to gauging the quality of labeled data. By furnishing researchers with uncomplicated yet illuminating insights, we empower them to make informed determinations concerning tag reliability. This, in turn, ushers in the creation of resilient and trustworthy machine learning models underpinned by meticulously labeled data.

5.2 Comparison between Manual Grading and Quality Control Strategies

In our research, we implemented four different strategies, and the next step involves comparing the scores obtained from these strategies with the manual grade scores assigned by the professor to each student. These manual grades by the professor scores served as the benchmark for evaluating the performance and effectiveness of our automated grading system. Each user received manual grades from the professor for their respective assignments, providing a reference point for assessing the accuracy and reliability of our automated scoring approach.

This comparison was essential to evaluate the effectiveness of each strategy in predicting the manual grades. To perform this comparison, we adopted an approach that started with using linear regression.

Linear regression is used to model the relationship between a dependent variable and

one or more independent variables. In this case, the dependent variable was the manual grade scores, and the independent variables were the scores obtained from each strategy (Krippendorff, fast tagging, pattern detection).

Before proceeding with the regression, we preprocessed the data by converting the 'assignment_id' and 'userid' columns from all the datasets (Krippendorff, fast tagging, pattern detection, and manual grading) into integers. This step was necessary to ensure consistency in the data types across all datasets, enabling a successful merging process.

After the preprocessing step, we merged the records from all datasets based on the 'assignment_id' and 'user_id' columns. This merging process allowed me to create a combined dataset that contained all the relevant information from each strategy alongside the manual grade scores. From this merged dataset, I obtained the datapoints representing a unique combination of 'assignment_id' and 'user_id.'

Initially, we conducted an analysis to assess the relationship between each independent feature and the target variable, which is the manual grade. We found that these variables displayed minimal correlation with the target. Consequently, opting for a linear regression model would likely yield unsatisfactory results.

In order to validate this observation, we established a null model and contrasted its Mean Squared Error (MSE) with that of the linear regression model. The outcome reaffirmed that the linear regression model inadequately predicts the target variable.

This lackluster performance could be attributed to a couple of factors:

Insufficient Data: It is plausible that the dataset's size is not sufficient to capture the nuances required for accurate predictions.

Lack of some parameters: Our manual grading process omitted the consideration of timing and the team's consensus during tagging. This omission might have introduced variability that the model struggles to accommodate.

In essence, the combination of weak variable correlations and these potential issues in the grading process could collectively contribute to the subpar performance of the linear regression model.

With the strategy scores obtained through linear regression, we then analyzed the correlation between the independent variables (alpha, repetition, IL_result) and the manual grade scores. The pandas correlation method provided us with correlation coefficients, indicating the strength and direction of the linear relationship between each independent variable and the manual grade scores.

Correlation coefficients range from -1 to 1, where -1 indicates a perfect negative correlation, 1 indicates a perfect positive correlation, and 0 indicates no linear correlation.

Positive correlation means that as one variable increases, the other tends to increase as well, and negative correlation means that as one variable increases, the other tends to decrease.

By analyzing these correlation coefficients, we gained insights into which independent variables (alpha, repetition, IL_result) had a stronger or weaker association with the manual grade scores. This analysis helped us to understand which aspects of the strategies had a more significant impact on predicting manual grades and could potentially guide further optimizations or improvements in the strategies.

In summary, the approach involved preprocessing the data, merging relevant records from different datasets, performing independent linear regression for each strategy to obtain strategy scores based on manual grading scores, and finally, analyzing the correlation between the independent variables and the manual grade scores. This comprehensive methodology allowed us to evaluate the effectiveness of each strategy and understand the factors contributing to the predictions made by these strategies.

The results obtained from the approach described above revealed a low correlation between the scores obtained from each strategy and the manual grades. There could be several reasons contributing to this finding.

Firstly, it was observed that manual grading primarily focused just on identifying patterns in student responses, and often overlooking timestamps and the level of agreement among grading teams. In contrast, the strategies employed in the research took into account a more comprehensive range of factors. By incorporating various parameters such as tagging speed, pattern detection, and agreement among teams, these strategies aimed to provide a more accurate and holistic evaluation of student performance. The greater emphasis on diverse aspects of student work allowed the strategies to offer a more nuanced assessment, which might not have been fully captured by the manual grading process.

Secondly, another probable cause for the low correlation between the strategy results and manual grades could be the limited availability of sufficient data that could be considered as a golden standard for the experiment. As the dataset only contained data from professors' grading, there might be the case where human emotions and fatigue affected the manual grading process itself. To address this, future plans involve increasing the dataset size by incorporating gradings from all Teaching Assistants (TAs). However, before considering TA gradings, it is crucial to assess their reliability.

To determine the reliability of TAs, a future plan involves comparing the comments given by each TA to the comments provided by the professor. This rigorous comparison will serve as a measure to gauge the consistency and accuracy of TAs' grading, ultimately contributing to the enhancement of the overall quality of crowd labeling. By ensuring

that the TAs' evaluations align closely with the professor's grading, the research team can establish a more trustworthy and expanded set of golden standard data, bolstering the reliability and validity of the entire annotation process.

To facilitate this comparative analysis between TA and professor grading, a strategic approach utilizing clustering techniques will be employed. Specifically, we envision clustering all the TAs who consistently provide similar grades when compared to the professor. This clustering will be based on an analysis of the comments provided by each TA, allowing us to identify patterns and similarities in their assessment approaches. By grouping TAs with similar grading tendencies, we can gain deeper insights into the commonalities and differences in their interpretations of assignments.

The application of clustering techniques offers a promising avenue for future research within the scope of our study. This approach not only enables us to identify patterns of agreement and divergence among TAs but also provides an opportunity to delve into the underlying factors that influence their grading decisions. Through this future endeavor, we aim to unravel the intricacies of TA evaluation, shedding light on potential areas of improvement and enhancing the overall robustness of the crowd labeling process.

Incorporating clustering as part of our comparative analysis adds an additional layer of sophistication to our quality control strategies. It aligns seamlessly with our broader objective of refining crowd labeling methodologies and underscores our commitment to continuously pushing the boundaries of quality assurance. As we embark on this next phase of investigation, we anticipate that the insights gleaned from clustering TAs' grading behaviors will significantly contribute to the ongoing evolution of crowd labeling practices, ensuring a high standard of accuracy and consistency in the annotations provided.

The Null Model often serves as a baseline, assuming the simplest possible model without any predictive features. The fact that the strategies' models perform similarly to the Null Model in terms of MSE indicates that there is room for improvement in the strategies' predictive power which can be enhanced by increasing the data size.

In conclusion, the approach revealed valuable insights into the effectiveness of the strategies in predicting manual grades. The low correlation with manual grading scores highlights the need to consider multiple factors and aspects of student work for a comprehensive evaluation. The future plan to incorporate TA gradings could provide a broader dataset, but assessing the reliability of TAs is essential to ensure data quality. By addressing these limitations and refining the strategies based on the findings, the research aims to enhance the accuracy and applicability of the evaluation process, ultimately benefiting the assessment of student performance.

Moreover, manual grading has its inherent limitations, which our automated approach effectively overcomes. Manual grading can be time-consuming and may lead to grading inconsistencies due to human subjectivity and fatigue. It is also challenging for professors to handle large volumes of assignments while maintaining consistency and objectivity in their grading decisions. Additionally, the manual process may not be able to efficiently analyze and consider a wide range of parameters and data points from each student's submission.

In contrast, our automated scoring system allows for a more efficient and objective evaluation process. By considering various factors, such as tagging speed, pattern detection, and agreement among teams, our approach provides a more comprehensive and accurate assessment of student performance. The automated nature of the system ensures consistent and reliable grading outcomes, reducing the potential for grading bias and variability.

Given the limitations of manual grading and the advantages offered by our automated approach, there is strong potential to replace manual grading with our system in certain educational settings. Implementing our automated scoring system can lead to significant time and resource savings, while also providing more detailed and insightful feedback to students. However, it is essential to acknowledge that no automated system is entirely flawless, and further validation and refinement may be necessary before full-scale adoption. As such, a hybrid approach, where manual grading and the automated scoring system complement each other, could be a prudent step towards optimizing the grading process in educational institutions.

It's essential to acknowledge that manual grading serves as a valuable benchmark, but its limitations necessitate the development of more sophisticated strategies like ours to achieve higher accuracy and reliability.

Despite the differences observed, our strategies demonstrated promising results in improving crowd labeling quality and providing reliable training data for machine learning models. By acknowledging the limitations of manual grading and leveraging the strengths of our strategies, we can lay the foundation for automated grading systems that streamline the assessment process while maintaining high data quality standards.

Through this comparison, we reinforce the significance of our research in addressing the challenges of crowd labeling and advancing the field of peer assessment-based educational methods. Moreover, we emphasize the potential of our strategies to revolutionize data annotation and improve machine learning model training in various domains beyond the educational context.

5.3 Limitations

In spite of the effectiveness of the developed strategies for quality control of crowd labeling, certain limitations should be considered:

Lack of comparison with existing approaches: One limitation of the quality control strategies is that there are not many of the previous validations or established benchmarks. Since these strategies may be a novel approach or one that hasn't been widely implemented before, there is limited knowledge regarding its effectiveness and validity and hence it becomes important to conduct thorough testing and validation to ensure the strategy's reliability and accuracy. This involves evaluating its impact on data quality, the consistency of results, and its compatibility with different datasets and labeling contexts.

Subjectivity in threshold determination: The determination of thresholds for the strategies, such as the minimum time threshold for fast tagging or the threshold for agreement/disagreement value, involves subjective decision-making. Selecting appropriate thresholds can be challenging and may require domain expertise or further experimentation to find the optimal values. Different threshold choices could potentially impact the performance and effectiveness of the strategies.

Limited evaluation on real-world data: While the strategies have been evaluated and validated on the available dataset, it is important to acknowledge the limitations of the dataset itself. The dataset used may not fully capture the complexities and nuances present in real-world crowd-labeling scenarios. Therefore, further validation and evaluation on diverse and larger-scale datasets are necessary to ascertain the robustness and generalizability of these strategies.

Limited generalizability: The strategies developed in this study may be specific to the dataset and context used. Different datasets or domains may exhibit variations in terms of labeling patterns, inter-rater agreement, and tagging behavior. Therefore, it is important to carefully evaluate the applicability and generalizability of these strategies to other datasets or tasks.

Reliance on assumptions: The strategies assume certain underlying assumptions, such as the assumption that adequate time spent on tagging implies a more thoughtful and accurate tagging process. However, these assumptions may not hold in all scenarios, as there could be instances where quick and accurate tagging is possible based on prior knowledge or expertise. Therefore, it is crucial to recognize the limitations and potential deviations from these assumptions in real-world scenarios. Another assumption of our fast tagging strategy is that all tags and tasks require the same time to complete. However, this may not

always be the case, as some tags may be more complex, or some tasks may require a more detailed analysis of the reviews and tag categories, which may take longer than the allotted time. Therefore, it is essential to consider the nature and complexity of the labeling task before implementing this strategy.

Acknowledging these limitations and addressing them in future research can contribute to the refinement and advancement of quality control strategies for crowd labeling.

5.4 Future Scope

Here are some potential areas for future research and improvement in the field of quality control for crowd labeling:

Hybrid approaches: Combining multiple quality control strategies can lead to more robust and reliable results. Future research can investigate the synergistic effects of combining different strategies, such as fast tagging, inter-rater reliability metrics, pattern detection, and machine learning-based approaches, to enhance the overall quality control process.

Domain-specific adaptation: The strategies developed in this study can be further tailored and adapted to specific domains or industry contexts. Different domains may have unique requirements and challenges regarding labeling quality and reliability. Future research can focus on customizing the strategies to specific domains, considering domain-specific characteristics, terminology, and labeling guidelines.

Improvement in labeling task: In addition to identifying low-performing workers, IRR can also be used to assess the overall quality of the labels. If the IRR is low, it may indicate that the labeling task is difficult or that the categories are not well-defined. This information can be used to improve the labeling task and to provide better guidance to the workers. IRR can also be used to identify the most reliable workers and to give them more responsibility in the labeling task.

Cross-platform applicability: Extending the applicability of the quality control strategies beyond a single online platform can be an interesting area for future research. Evaluating the effectiveness of the strategies across different crowd-labeling platforms or even in offline settings can provide insights into their generalizability and practicality in diverse contexts.

User feedback integration: Incorporate user feedback and preferences into the quality control process. Allow users, such as domain experts or end-users of the labeled data, to provide feedback on the quality and relevance of the labels, enabling iterative improvements

and fine-tuning of the quality control strategies.

By addressing these future scope areas, researchers can advance the field of quality control in crowd labeling, enhancing the reliability, efficiency, and scalability of the labeling process and the quality of the resulting labeled datasets.

CHAPTER

6

CONCLUSIONS

In the realm of data annotation and machine learning, the role of crowd labeling has grown substantially, offering a cost-effective and scalable approach to data tagging and annotation. However, the inherent challenges of maintaining data quality, ensuring reliability, and mitigating biases have become increasingly evident. This thesis embarked on a journey to address these challenges head-on and proposed innovative strategies to enhance the quality and credibility of crowd labeling, contributing to the advancement of automated grading systems and machine learning model training.

Through a meticulous exploration of the landscape of crowd labeling, we identified the limitations and obstacles that hindered its potential. The need to ensure accurate, consistent, and unbiased annotations emerged as a paramount concern. To tackle these issues, we devised and meticulously executed four distinct strategies, each targeting a specific facet of quality control. These strategies included fast tagging detection, inter-rater reliability assessment, pattern detection, and agreement/disagreement analysis.

The empirical validation of these strategies showcased promising results. The fast tagging detection strategy unearthed students who rushed through tagging tasks, often compromising data quality. It provides researchers with a quick and easy way to identify potential quality issues and intervene early to address them. However, researchers should be

cautious and ensure that the quality of the assigned labels is not compromised by focusing solely on the speed of tagging.

Our inter-rater reliability analysis revealed the significance of agreement among annotators, shedding light on instances where diverse interpretations impacted labeling consistency. Our results demonstrate that researchers can set a threshold value to classify users as reliable or unreliable based on their judgment of what constitutes a reliable level of agreement.

The pattern detection approach unmasked students who were seemingly following specific patterns, reinforcing the importance of vigilant evaluation. Our approach utilized the Progressive Timelist-Based Verification algorithm, which allowed us to identify patterns that repeated a minimum number of times within a specified range of pattern lengths.

Lastly, the agreement/disagreement analysis provided a mechanism to filter out unreliable annotations through consensus-based decision-making. The IRR agreement/disagreement strategy evaluates the level of agreement or disagreement for each tag based on the mode value. By calculating IRR values using the mode metric, we were able to assess the degree of agreement between the crowd workers and identify tags with high levels of reliability.

The results of our quality control research paper show a comparative analysis between the metrics obtained from our study and the manual grading conducted by the professor for the same user and assignment. The comparative analysis of our strategies with manual grading underscored their efficacy in providing a more comprehensive and nuanced evaluation of student performance. While manual grading serves as a conventional benchmark, our strategies leverage additional parameters and contextual insights, ultimately contributing to a more holistic assessment.

To begin the analysis, we collected the metric results generated by each of our strategies for each user in the assignment. These metrics included pattern identification and repetition count, fast tagging timestamp difference values, Krippendorff's alpha values, and Agreement /Disagreement fractional values of tags for the relevant data points and users. These metrics provide quantitative data on the performance of each user in completing the assignment. Simultaneously, the professor performed manual grading for the same users and assignments, providing a benchmark for comparison. Our research offers a compelling argument for adopting our study's approach as a reliable and efficient alternative to manual grading tasks. The potential benefits include increased efficiency, reduced subjectivity, and standardized evaluation processes.

The evidence we have supports the justification for adopting our study's approach as a potential replacement for manual grading tasks in crowd labeling, leading to improved

efficiency, accuracy, and cost-effectiveness in various domains relying on crowd-sourced data.

By implementing these proposed strategies, researchers and practitioners can improve the accuracy, reliability, and consistency of labeled datasets, which in turn enhances the effectiveness of downstream tasks such as machine learning model training and evaluation. While these strategies have shown promising results, it is important to acknowledge their limitations and further explore their applicability across different domains and datasets. Our research does acknowledge certain limitations, such as the dependence on available data, the necessity for appropriate threshold setting, and the inherent subjectivity in some aspects of quality control. However, these limitations open doors to further exploration and refinement, highlighting the evolving nature of crowd labeling and quality assurance.

The significance of this work extends beyond the academic realm. The proposed strategies hold the potential to revolutionize the way data is annotated, impacting fields ranging from education to healthcare and beyond. Automated grading systems stand to benefit immensely from the enriched training datasets generated through our strategies, potentially reducing manual effort and enhancing model performance.

As we culminate this journey, it is evident that the path to reliable and high-quality crowd labeling is multifaceted. The strategies presented in this thesis mark a decisive step towards realizing this goal. By establishing a framework for robust quality control, we strive to empower researchers, educators, and data scientists to harness the true potential of crowd labeling. Through collaboration, open research, and continuous refinement, we envision a future where crowd labeling becomes an invaluable asset in data annotation, model training, and knowledge dissemination.

The results of our research contribute to the field of crowd labeling quality control by providing a robust and objective approach to identifying unreliable workers. By removing these unreliable contributions, researchers can improve the overall accuracy and reliability of crowd labels, leading to better outcomes and more informed decisions. Overall, this research contributes to the advancement of quality control techniques in crowd labeling, enabling improved data reliability, increased efficiency, and enhanced performance in various domains relying on crowd-sourced labeled datasets.

In closing, this thesis heralds a new era in the realm of crowd labeling, instilling confidence in the annotations produced and fostering a culture of excellence in data-driven research and applications. As we embark on the next phase of technological evolution, let this work stand as a testament to the power of innovation, collaboration, and dedicated pursuit of quality.

REFERENCES

[1] Topping, K. J. (2009). Peer assessment. Theory into Practice, 48(1), 20-27.

[2] Boud, D., Cohen, R., Sampson, J. (2001). Peer learning and assessment. Assessment Evaluation in Higher Education, 26(3), 253-266.

[3] Falchikov, N. (2005). Improving assessment through student involvement: Practical solutions for aiding learning in higher and further education. Routledge.

[4] Johnson, D. W., Johnson, R. T., Smith, K. A. (1991). Cooperative learning: Increasing college faculty instructional productivity (ASHE-ERIC Higher Education Report No. 4). ERIC Clearinghouse on Higher Education

[5] Snow, R., O'Connor, B., Jurafsky, D., Ng, A. Y. (2008). Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 254-263.

[6] Karger, D. R., Oh, S., Shah, D. (2011). Iterative learning for reliable crowdsourcing systems. Proceedings of the Conference on Neural Information Processing Systems, 1953-1961.

[7] Balahur, A., Turchi, M., Steinberger, R. (2013). Improving sentiment analysis in an under-resourced language using crowdsourcing. Information Processing Management, 49(2), 407-418. https://doi.org/10.48550/arXiv.1309.6202

[8] Welinder, P., Branson, S., Belongie, S., Perona, P. (2010). The multidimensional wisdom of crowds. Proceedings of the Conference on Computer Vision and Pattern Recognition, 2424-2431

[9] Karrie Karahalios, Andrés Monroy-Hernández, Airi Lampinen, and Geraldine Fitzpatrick (Eds.). 2018. Proc. ACM Hum.-Comput. Interact. 2, CSCW (November 2018).

[10] Heim, Eric. Large-scale medical image annotation with quality-controlled crowdsourcing. Diss. 2018.

[11] Crowd-sourcing Annotation of Clinical Texts: Measuring Annotator Engagement and Agreement" by Pradhan, et al. (2014)

[12] Michael S. Bernstein, Joel Brandt, Robert C. Miller, and David R. Karger. 2011. Crowds in two seconds: enabling realtime crowd-powered interfaces. In Proceedings of the 24th annual ACM symposium on User interface software and technology (UIST '11). Association for Computing Machinery, New York, NY, USA, 33–42. https://doi.org/10.1145/2047196.2047201

[13] Som, Anirudh, et al. "Automated Student Group Collaboration Assessment and Recommendation System Using Individual Role and Behavioral Cues." Frontiers in Computer Science 3 (2021): 728801.

[14] Lease, Matthew. "On quality control and machine learning in crowdsourcing." Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence. 2011.

[15] Gehringer, E. F. (2012). Expertiza: A Web-Based Peer Review System. IEEE Transactions on Education, 55(4), 468-475. doi: 10.1109/TE.2012.2185043

[16] Gehringer, E. F. (2010). Peer Review in a Web 2.0 World: Using a Web-Based Peer Review System to Improve Student Writing. IEEE Transactions on Professional Communication, 53(3), 278-292. doi: 10.1109/TPC.2010.2047008

[17] Gehringer, E. F. (2011). Expertiza: A Peer Review System that Improves Student Learning. In Proceedings of the 41st ACM Technical Symposium on Computer Science Education (pp. 261-266). doi: 10.1145/1953163.1953251

[18] Edward F. Gehringer, Luke M. Ehresman, and Dale J. Skrien. 2006. Expertiza: students helping to write an OOD text. In Companion to the 21st ACM SIGPLAN symposium on Object-oriented programming systems, languages, and applications (OOPSLA '06). Association for Computing Machinery, New York, NY, USA, 901–906. https://doi.org/10.1145/1176617.1176742

[19] Felix Garcia-Loro, Sergio Martin, José A. Ruipérez-Valiente, Elio Sancristobal, Manuel Castro,Reviewing and analyzing peer review Inter-Rater Reliability in a MOOC platform, Computers Education, Volume 154, 2020, 103894, ISSN 0360-1315, https://doi.org/10.1016/j.compedu.2020.103894. (https://www.sciencedirect.com/science/article/pii/S0360131520300932)

[20] Krippendorff, K. (2011). Computing Krippendorff's Alpha-Reliability. Retrieved from https://repository.upenn.edu/asc_papers/43

[21] Felix Garcia-Loro, Sergio Martin, José A. Ruipérez-Valiente, Elio Sancristobal, Manuel Castro, Reviewing and analyzing peer review Inter-Rater Reliability in a MOOC platform, Computers and Education, Volume 154, 2020, 103894, ISSN 0360-1315, https://doi.org/10.1016/j.compedu.2020.103894. (https://www.sciencedirect.com/science/article/pii/S0360131520300932)

[22] Zapf, A., Castell, S., Morawietz, L. et al. Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate?. BMC Med Res Methodol 16, 93 (2016). https://doi.org/10.1186/s12874-016-0200-9

[23] Scott WA. Reliability of content analysis: the case of nominal scale coding. Public Opinion Quarterly. 1955;XIX:321–5.

[24] Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20:37–46.

[25] Hallgren KA. Computing inter-rater reliability for observational data:_an overview and tutorial. Tutor Quant Methods Psychol. 2012;8(1):23–4.

[26] Thompson WD, Walter SD. A reappraisal of the kappa coefficient. J Clin Epidemiol. 1988;41(10):949–58.

[27] Kraemer HC, Bloch DA. Kappa coefficients in epidemiology: an appraisal of a reappraisal. J Clin Epidemiol. 1988;41(10):959–68.

[28] Feinstein A, Cicchetti D. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol. 1990;43(6):543–9.

[29] Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull. 1971;76:378–82.

[30] Siegel S, Castellan Jr NJ. Nonparametric Statistics for the Behavioral Sciences. 2nd ed. New York: McGraw-Hill; 1988.

[31] Krippendorff K. Estimating the reliability, systematic error, and random error of interval data. Educ Psychol Meas. 1970;30:61–70.

[32] Y. Zhang and E. F. Gehringer, "Can Students Produce Effective Training Data to Improve Formative Feedback?," 2021 IEEE Frontiers in Education Conference (FIE), Lincoln, NE, USA, 2021, pp. 1-7, doi: 10.1109/FIE49875.2021.9637414.

[33] Xueying Zhan, Yaowei Wang, Yanghui Rao, and Qing Li. 2019. Learning from Multi-annotator Data: A Noise-aware Classification Framework. ACM Trans. Inf. Syst. 37, 2, Article 26 (April 2019), 28 pages. https://doi.org/10.1145/3309543

[34] Mason, W., Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. Behavior Research Methods, 44(1), 1-23.

[35] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 2018. Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions. ACM Comput. Surv. 51, 1, Article 7 (January 2019), 40 pages. https://doi.org/10.1145/3148148

[36] Upchurch, P., Sedra, D., Mullen, A., Hirsh, H., Bala, K. (2016). Interactive Consensus Agreement Games for Labeling Images. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 4(1), 239-248. https://doi.org/10.1609/hcomp.v4i1.13293

[37] Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd work. In Proceedings of the 2013 conference on Computer supported cooperative work (CSCW '13). Association for Computing Machinery, New York, NY, USA, 1301–1318. https://doi.org/10.1145/2441776.2441923

[38] Snow, R., O'Connor, B., Jurafsky, D., Ng, A. Y. (2008). Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 254-263.

[39] Kulesza, T., Burnett, M., Wong, W. K., Stumpf, S., Perona, S., Ko, A. J. (2014). Principles of explanatory debugging to personalize interactive machine learning. Proceedings of the 2014 Conference on Intelligent User Interfaces, 126-137.

[40] Panagiotis G. Ipeirotis, Foster Provost, and Jing Wang. 2010. Quality management on Amazon Mechanical Turk. In Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP '10). Association for Computing Machinery, New York, NY, USA, 64–67. https://doi.org/10.1145/1837885.1837906

[41] Kazai, Gabriella, and Imed Zitouni. "Quality management in crowdsourcing using gold judges behavior." Proceedings of the Ninth ACM International Conference on Web Search and Data Mining. 2016.

[42] Huang, Kuo-Yu, and Chia-Hui Chang. "Mining periodic patterns in sequence data." Data Warehousing and Knowledge Discovery: 6th International Conference, DaWaK 2004, Zaragoza, Spain, September 1-3, 2004. Proceedings 6. Springer Berlin Heidelberg, 2004.