Marie-Elise Latorre

Student Number : 260981230

Final Project

Aggregation of Crowdsourced Labels

McGill University

Due Date : December 23, 2024

COMP 598 - Natural Computation

Introduction:
Crowdsourcing has emerged as a powerful tool for tackling large-scale tasks that require human judgment, such as document relevance assessment, image classification, and language translation (Beck et al.). Platforms like Amazon Mechanical Turk (MTurk) have enabled researchers to collect vast amounts of labeled data from diverse contributors, making it possible to perform tasks that were previously infeasible due to resource constraints (Newman). However, the quality of crowdsourced labels can vary significantly due to differences in worker reliability, task difficulty, and subjective interpretations of the task requirements. Aggregating these noisy and inconsistent labels into high-quality, reliable outputs is a critical challenge for researchers and practitioners alike ("Aggregation Methods").

This paper focuses on the task of aggregating crowdsourced relevance judgments for English web pages into a single consensus label for each document-query pair. The dataset, drawn from the 2010 TREC Relevance Feedback Track, includes 98,453 relevance judgments from 766 MTurk workers across 20,232 tasks. To address the challenges of label noise and worker variability, I implemented and evaluated seven aggregation methods, ranging from simple statistical techniques to advanced machine learning models. Early experiments focused on majority voting and weighted voting schemes that leveraged worker accuracy. These methods provided valuable insights into the limitations of simple aggregation strategies and the importance of considering worker reliability. Subsequent experiments explored machine learning-based approaches, incorporating task-level and worker-level features to train predictive models. By systematically refining these methods, I sought to address key challenges such as class imbalance, task ambiguity, and worker specialization.

This paper aims to answer the following research questions: How can we effectively aggregate noisy crowdsourced labels to maximize accuracy? What features and methods best capture the variability in worker performance and task characteristics? How do advanced machine learning models compare to traditional aggregation techniques in this context? By answering these questions, this study provides practical insights into the design of effective label aggregation systems and contributes to the broader understanding of crowdsourcing as a tool for large-scale human computation.

Dataset Exploration:
I began the analysis by examining the structure and content of the dataset, which consists of 98,453 relevance judgments provided by 766 Mechanical Turk workers across 20,232 unique tasks. The tasks involved evaluating the relevance of English Web pages from the ClueWeb09 collection in response to search queries taken from the TREC 2009 Million Query track, a standard resource for web search evaluation. Each entry includes the following key attributes: a topic identifier (topicID), a unique worker identifier (workerID), a document identifier (docID), a gold label (gold), and the worker-provided label (label). Each task corresponds to a pair of a

search topic and a document, and workers were asked to evaluate the relevance of these documents using a ternary scale: highly relevant (2), relevant (1), non-relevant (0), with an additional category for broken links (-2). Of the total tasks, 26.6% (3,277) have gold labels—ground truth judgments provided by NIST—which serve as a benchmark for assessing worker performance."

To explore worker participation, I calculated the total number of judgments each worker contributed. The distribution revealed that while a few workers contributed a large number of judgments (up to 7,920 tasks for the most active worker), the majority of workers provided far fewer judgments. As shown in Figure 1, most workers completed fewer than 500 tasks, highlighting a common imbalance in crowdsourced data, where contributions are dominated by a small subset of highly active workers.

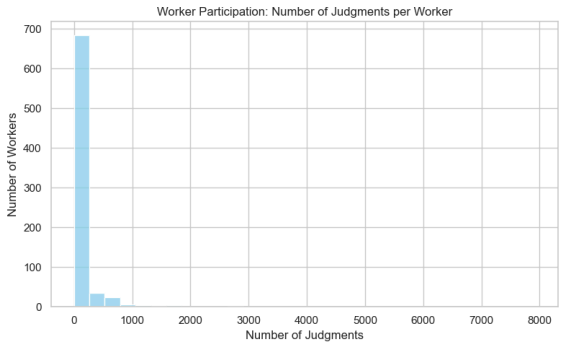

Figure 1: Worker Participation Distribution *– Histogram showing the number of judgments completed by each worker.*

Next, I analyzed task coverage by examining how many workers judged each task. As depicted in Figure 2, most tasks were evaluated by exactly five workers, reflecting a design choice to achieve redundancy and consensus and mitigating errors from individual workers.

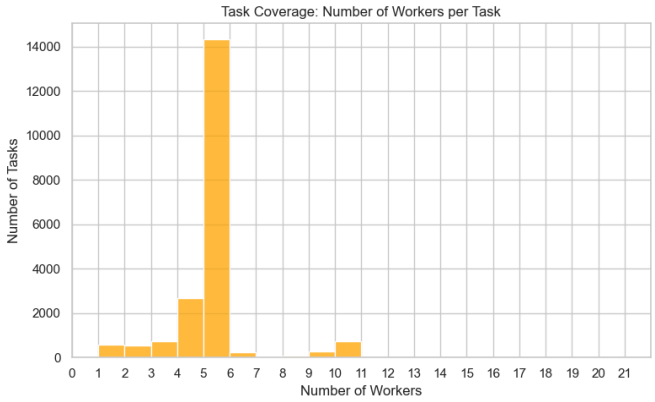

Figure 2: Task Coverage Distribution*– Histogram displaying the number of workers who judged each task.*

The gold labels distribution (Figure 3) shows that non-relevant (0) pages dominate the dataset, followed by a roughly equal number of tasks labeled as relevant (1) and highly relevant (2). A significant portion of tasks were also marked as broken links (-2), further validating the dataset's inclusion of intentionally non-existent URLs as a quality assurance measure.

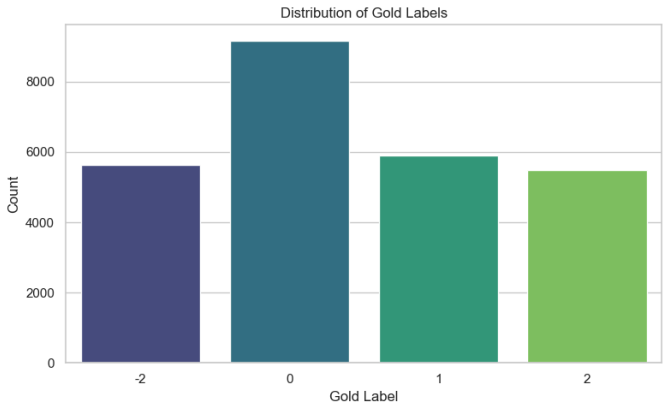

Figure 3: Distribution of Gold Labels *– Bar chart illustrating the frequency of gold labels across all tasks.*

To compare worker-provided labels to gold labels, I focused on tasks with gold labels (Figure 4) (so no unknown (-1)). This comparison revealed important patterns: workers generally agreed on identifying broken links, with most gold-labeled -2 tasks correctly identified by workers. However, for tasks labeled as relevant (1) or highly relevant (2), worker judgments exhibited higher variability. Workers frequently misclassified relevant pages as non-relevant (0) or struggled to distinguish between 1 and 2. This suggests potential challenges in subjective judgments of relevance, where clear distinctions between relevance levels may not always be apparent.

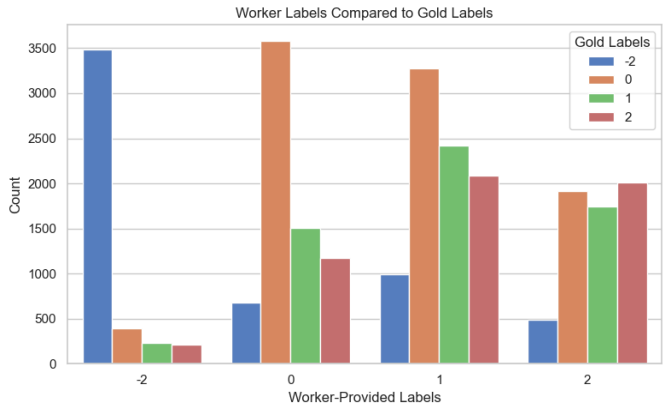

Figure 4: Worker Labels Compared to Gold Labels *– Comparison of worker-provided labels against gold-standard labels for tasks with ground truth.*

Methods and Results:

To explore the optimal approach for aggregating crowdsourced relevance judgments into high-quality labels, I conducted a series of seven experiments. These experiments evolved from basic statistical aggregation techniques to advanced machine learning (ML) models, each incorporating progressively refined strategies to address issues such as worker reliability, task ambiguity, and class imbalance. The experiments also generated visual insights in the form of confusion matrices (Figures 5–11) to evaluate performance across all task categories.

The first experiment implemented a majority voting method as a baseline (Morton et al.). In this approach, the label with the highest number of worker votes was selected as the final consensus label for each task. In cases of ties, a random label among the most frequent was chosen. This simple method treated all workers equally and ignored individual reliability or performance. The majority voting approach achieved an accuracy of 54.30%. The confusion matrix for this method (Figure 5) reveals that the model performed reasonably well for identifying broken links (-2), correctly classifying 939 tasks. However, significant misclassifications were observed across the remaining categories, particularly between adjacent labels such as 0 (non-relevant), 1 (relevant), and 2 (highly relevant). The results highlighted the limitations of majority voting, as it failed to account for differences in worker quality, leading to substantial noise in the aggregated labels.

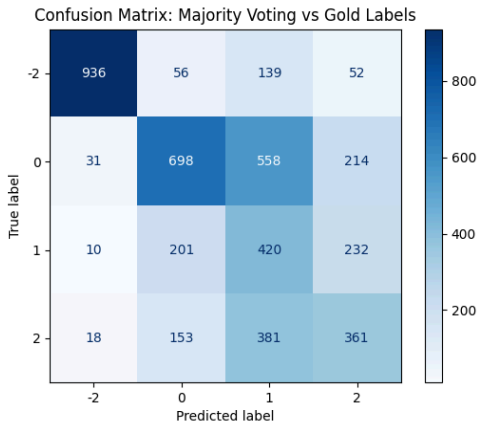

Figure 5: Confusion Matrix for Majority Voting – A visual representation of model predictions compared to gold labels for the majority voting approach, highlighting misclassifications across task categories.

To address this limitation, the second experiment introduced weighted voting, where workers' contributions were weighted based on their accuracy on gold-labeled tasks (Meyen et al.). Each worker's accuracy was computed as the proportion of their judgments matching the gold labels. Tasks were then aggregated using these worker-specific weights to favor judgments from more reliable workers. This weighted voting method improved the overall accuracy to 62.13%. The confusion matrix (Figure 6) demonstrated a notable reduction in misclassifications, particularly

for broken links and non-relevant tasks. However, the method continued to struggle with finer distinctions between relevance levels, reflecting the subjectivity of these task categories.

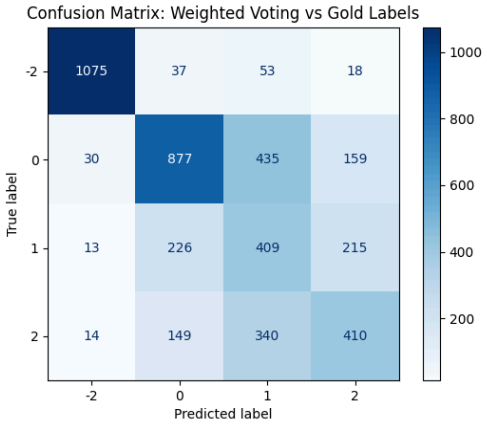

Figure 6: Confusion Matrix for Weighted Voting – Comparison of model predictions versus gold labels for weighted voting, where worker accuracy was used to weight contributions.

In the third experiment, I refined the weighted voting approach by introducing an accuracy threshold. Workers whose accuracy on gold-labeled tasks fell below a specified threshold (60%) were excluded from contributing to the aggregated labels. Additionally, the weights of reliable workers were squared to further amplify their influence. This method aimed to filter out unreliable workers and prioritize the judgments of high-performing contributors. After filtering, 332 workers remained in the aggregation process. The resulting accuracy was 61.32%, slightly lower than the previous weighted voting approach. The confusion matrix (Figure 7) showed improvements for non-relevant (0) and broken link tasks (-2) but revealed ongoing difficulties with adjacent relevance categories. These results suggest that while thresholding effectively removes noise, it may reduce diversity in label contributions, especially for tasks with limited worker participation.

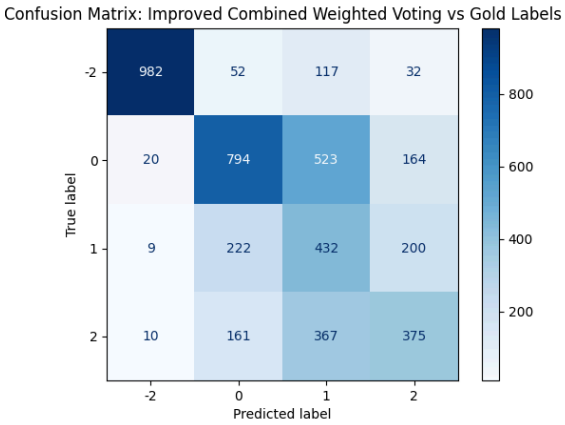

Figure 7: Confusion Matrix for Improved Weighted Voting with Thresholding – Results of weighted voting after filtering workers below a 60% accuracy threshold, showing the impact on task categories.

Recognizing the potential for workers to perform differently across task types, the fourth experiment implemented task-specific weighted voting. In this method, workers' accuracy was calculated separately for each task type: broken links, non-relevant, relevant, and highly relevant documents. Workers were then assigned task-specific weights based on their performance for the corresponding label category. This approach aimed to capture worker specialization, ensuring that their contributions were valued more highly for tasks they excel at. Task-specific weighted voting resulted in a significant accuracy improvement, reaching 80.04%. The confusion matrix (Figure 8) revealed strong performance for all categories, particularly for broken links and non-relevant documents. Misclassifications between relevant (1) and highly relevant (2) tasks were reduced but not fully eliminated. The results highlight the value of dynamically weighting workers based on their expertise for specific task types.

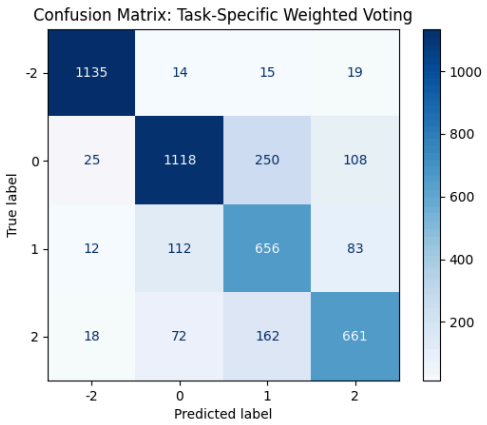

Figure 8: Confusion Matrix for Task-Specific Weighted Voting – Performance of task-specific weighted voting, where worker contributions were dynamically weighted based on their accuracy for specific task types.

The limitations of statistical aggregation methods motivated a shift toward machine learning-based approaches in the final three experiments. In Experiment 5, I trained a Random Forest classifier to predict gold labels using a rich set of task-level and worker-level features (Jason Brownlee). Task-level features included counts of worker-provided labels, worker experience (mean and max), and worker accuracy (mean and max). By incorporating these features, the model learned to identify patterns in worker judgments and their relationship to true labels. The Random Forest classifier achieved an accuracy of 83.40%, far surpassing earlier methods. The confusion matrix (Figure 9) showed near-perfect performance for broken links and notable improvements across relevance categories. This result demonstrated the power of supervised learning to extract insights from worker data and task characteristics.

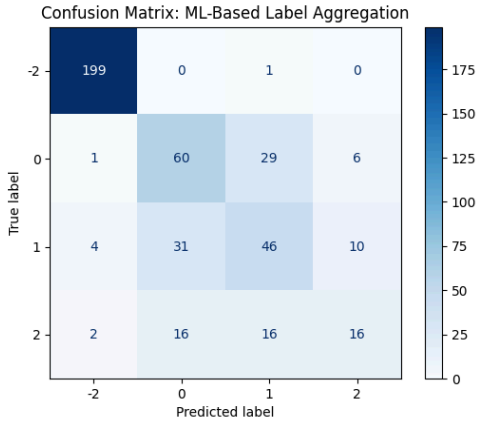

Figure 9: Confusion Matrix for Random Forest Classifier – Aggregation of labels using a Random Forest model trained on worker- and task-level features, with predictions compared to gold labels.

Building on the success of the Random Forest model, Experiment 6 introduced XGBoost, an optimized gradient-boosting classifier, to further enhance performance (GeeksforGeeks). Additional features, such as label entropy (a measure of worker disagreement), were included to provide the model with richer information about task ambiguity. Hyperparameter tuning using GridSearchCV ensured that the model was optimized for performance. The XGBoost classifier achieved an accuracy of 85.84%, improving on the Random Forest results. The confusion matrix (Figure 10) revealed fewer misclassifications between relevance categories, particularly for tasks labeled as 0 (non-relevant) and 1 (relevant). The inclusion of label entropy helped the model better handle tasks with high disagreement among workers, further refining the predictions.

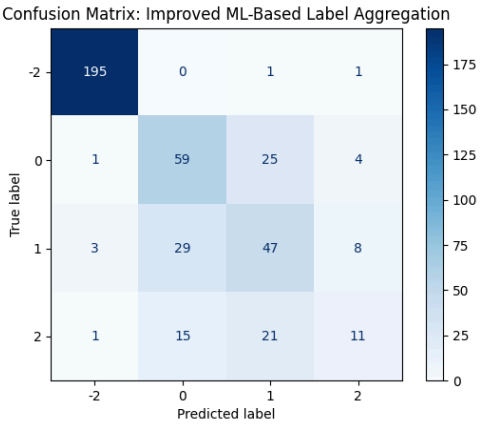

Figure 10: Confusion Matrix for XGBoost Model – Results of the XGBoost classifier trained with optimized hyperparameters and additional features, such as label entropy, showing further improvements in label aggregation.

In the seventh and final experiment, I addressed the issue of class imbalance by incorporating class weights into the XGBoost model. Since categories like broken links (-2) and highly relevant documents (2) were underrepresented in the dataset, class weights were calculated to ensure that these classes were given higher importance during training. The class-balanced XGBoost model achieved the highest accuracy of 85.96%, marking the peak performance across all experiments. The confusion matrix (Figure 11) demonstrated near-perfect predictions for broken links, with further reductions in misclassifications for the remaining categories. While the gains were modest compared to the previous XGBoost model, the introduction of class weights ensured more equitable performance across all task types, particularly for underrepresented labels.

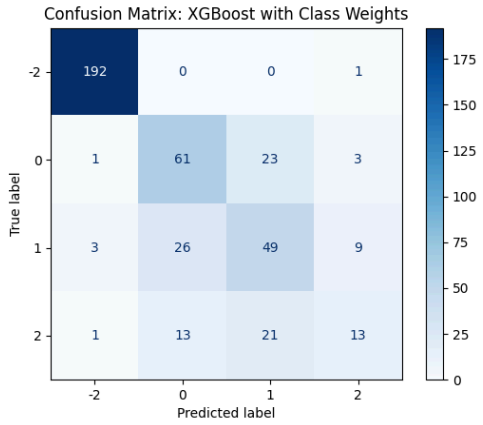

Figure 11: Confusion Matrix for XGBoost with Class Weights – Performance of the XGBoost model with class weights to address class imbalance, achieving the highest overall accuracy.

Across the seven experiments, the results demonstrate a clear progression in accuracy and reliability as the methods evolved from simple statistical aggregation to advanced machine learning approaches. Majority voting served as a baseline, achieving moderate performance but suffering from noise due to unreliable workers. Weighted voting methods improved performance by incorporating worker accuracy, with task-specific weights yielding significant gains. The introduction of machine learning marked a turning point, with the Random Forest and XGBoost models outperforming all previous methods by leveraging richer task- and worker-level features. Finally, addressing class imbalance through class weights further refined the XGBoost model, achieving an accuracy of 85.96%.

Discussion:

The progression of experiments in this study highlights the challenges and opportunities in aggregating crowdsourced labels for relevance judgments. By evolving from basic statistical methods to advanced machine learning approaches, we systematically addressed issues such as worker reliability, task ambiguity, and class imbalance, achieving significant improvements in accuracy and label quality. This section reflects on the key findings, the implications of the methods used, and areas for future exploration.

One of the fundamental challenges in crowdsourced data aggregation is accounting for worker variability. Early experiments, such as majority voting, failed to address this, resulting in noisy and unreliable aggregated labels. Weighted voting methods, which incorporated worker accuracy as a weighting factor, demonstrated the importance of recognizing individual worker contributions. However, the results also showed that worker performance varies not only across tasks but also across task types. For instance, task-specific weighted voting revealed that some workers excel at identifying broken links, while others perform better on relevance judgments. This underscores the need for aggregation methods that dynamically adjust to worker specialization, a theme echoed in the significant performance gains observed in Experiment 4.

Task characteristics also played a crucial role in the model's performance. Certain task categories, such as broken links (-2), consistently achieved high accuracy across methods, reflecting their unambiguous nature. In contrast, relevance judgments (0, 1, and 2) proved more subjective, with frequent misclassifications between adjacent categories. This suggests that the inherent ambiguity of these tasks, rather than model limitations, may be a persistent barrier to achieving perfect accuracy.

The transition to machine learning methods marked a major turning point in the study. Both the Random Forest model (Experiment 5) and the XGBoost model (Experiment 6) demonstrated the ability to leverage rich feature sets to make more informed predictions. Features such as worker accuracy, experience, label distributions, and label entropy allowed the models to capture nuanced relationships between worker judgments and true labels. These methods achieved significant accuracy improvements, with XGBoost outperforming all previous approaches.

The success of the machine learning models highlights the importance of feature engineering in crowdsourced data aggregation. For instance, label entropy—a measure of worker disagreement—proved particularly valuable in distinguishing between tasks with high variability in judgments. This suggests that future work could explore additional features, such as task difficulty or worker consistency, to further enhance model performance.

Despite their strengths, machine learning models require careful consideration of computational complexity and interpretability. While methods like Random Forest and XGBoost are well-suited for structured data, their reliance on hyperparameter tuning and large datasets may pose challenges for real-time applications. Additionally, although these models performed well overall, their confusion matrices (Figures 5–11) highlight persistent struggles with relevance categories, indicating room for further refinement.

The introduction of class weights in Experiment 7 demonstrated the importance of addressing class imbalance in the dataset. Broken links and highly relevant documents were underrepresented, which could have biased the model toward more frequent labels. By calculating and incorporating class weights, the XGBoost model achieved the highest accuracy of 85.96%, improving predictions for underrepresented categories. However, the modest performance gain suggests that class weighting alone may not fully resolve the challenges posed by imbalanced data.

This finding underscores the need for a multi-faceted approach to address class imbalance. In addition to class weights, techniques such as oversampling minority classes, synthetic data generation, or incorporating external domain knowledge could help improve performance. Moreover, class imbalance may exacerbate misclassifications for ambiguous relevance categories, highlighting the need for models that explicitly account for uncertainty in worker judgments.

Despite the advances achieved in this study, some challenges remain. Misclassifications between relevance categories (0, 1, and 2) were evident across all methods, including the final XGBoost model with class weights. These errors are likely rooted in the subjectivity of relevance judgments, as workers may interpret the relevance scale differently or struggle with borderline cases. While task-specific weighting and label entropy helped mitigate these issues, they did not fully resolve them, suggesting that further improvements may require more sophisticated approaches.

Another persistent challenge lies in the diversity of worker behavior. While accuracy-based weighting methods successfully prioritized reliable workers, they may inadvertently exclude contributions from less experienced but highly capable participants. Future methods could

explore adaptive weighting schemes that account for both long-term reliability and short-term task performance, ensuring that valuable input is not overlooked.

The findings of this study have broader implications for the design of crowdsourcing systems and label aggregation methods. The transition from simple aggregation to machine learning highlights the potential of data-driven approaches to improve the quality of crowdsourced outputs. By leveraging worker-level and task-level features, these methods can adapt to the inherent variability of human contributions, producing labels that align more closely with ground truth.

Looking ahead, future research could explore ensemble methods that combine the strengths of multiple aggregation strategies, such as weighted voting and machine learning. Additionally, incorporating unsupervised techniques to infer missing gold labels could expand the applicability of these methods to datasets without ground truth. Finally, ethical considerations, such as ensuring fair compensation for workers and addressing potential biases in task design, should remain central to the development of crowdsourcing systems.

Conclusion:

The results of this study demonstrate that effective label aggregation requires a balance between simplicity, adaptability, and computational complexity. While majority voting provides a straightforward baseline, more advanced methods—particularly machine learning—are essential for achieving high accuracy in noisy and subjective tasks. By incorporating features such as worker specialization, task ambiguity, and class balancing, the XGBoost model achieved the highest accuracy of 85.96%, setting a new benchmark for crowdsourced relevance judgments. These findings provide a foundation for future innovations in crowdsourcing, highlighting the potential of hybrid models and adaptive strategies to address the complexities of human computation.

References:

Beck, Susanne, et al. “Crowdsourcing Research Questions in Science.” *Research Policy*, vol. 51, no. 4, May 2022, p. 104491, https://doi.org/10.1016/j.respol.2022.104491.

Newman, Andy. “I Found Work on an Amazon Website. I Made 97 Cents an Hour.” *The New York Times*, 15 Nov. 2019, www.nytimes.com/interactive/2019/11/15/nyregion/amazon-mechanical-turk.html.

“Aggregation Methods.” *Aggregation Methods*, 2024, toloka.ai/knowledgebase/aggregation/. Accessed 18 Dec. 2024.

Morton, Rebecca B., et al. “The Dark Side of the Vote: Biased Voters, Social Information, and Information Aggregation through Majority Voting.” *Games and Economic Behavior*, vol. 113, Jan. 2019, pp. 461–481, https://doi.org/10.1016/j.geb.2018.10.008. Accessed 9 Nov. 2020.

Meyen, Sascha, et al. “Group Decisions Based on Confidence Weighted Majority Voting.” *Cognitive Research: Principles and Implications*, vol. 6, no. 1, 15 Mar. 2021, media.proquest.com/media/hms/PFT/1/fRbVI?_s=RjutwF3LDWbSN4o2jFl9F57Diwc%3D, https://doi.org/10.1186/s41235-021-00279-0. Accessed 24 May 2021.

Jason Brownlee. “Bagging and Random Forest Ensemble Algorithms for Machine Learning.” *Machine Learning Mastery*, 2 June 2019, machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithms-for-machine-learning/.

GeeksforGeeks. “XGBoost.” *GeeksforGeeks*, 18 Sept. 2021, www.geeksforgeeks.org/xgboost/.