Journal of Machine Learning Research 133:156–169, 2021 NeurIPS 2020 Competition and Demonstration Track

Semi-Automated Data Labeling

Michael Desmond MDESMOND@US.IBM.COM
Evelyn Duesterwald DUESTER@US.IBM.COM
Kristina Brimijoin KBRIMIJ@US.IBM.COM
Michelle Brachman MICHELLE.BRACHMAN@IBM.COM
Qian Pan QIAN.PAN@IBM.COM
IBM Thomas J Watson Research Center,
1101 Kitchawan Rd,
Yorktown Heights,
New York, USA.

Editors: Hugo Jair Escalante and Katja Hofmann

Abstract

Labeling data is often a tedious and error-prone activity. However, organizing the labeling experience as a human-machine collaboration has the potential to improve label quality and reduce human effort. In this paper we describe a semi-automated data labeling system which employs a predictive model to guide and assist the human labeler. The model learns by observing labeling decisions, and is used to recommend labels and automate basic functions in the labeling interface. Agreement between the labeler and the model is tracked and presented via a system of checkpoints. At each checkpoint the labeler has the opportunity to delegate the remainder of the labeling task to the model.

Keywords: Data Labeling, Human Computer Interaction, Interactive Machine Learning

1. Introduction

Labeling data is an important but tedious activity that is necessary during the development of supervised machine learning systems. The nature of data means that there is generally a spectrum of difficulty involved with any labeling task. Some instances will be ambiguous and difficult to label, and may result in mislabeled data. On the other hand more obvious instances may cause labelers to become bored with the task.

Desmond et al. (2021) studied the effects of AI assistance on labeling performance, and found that showing labelers the labels predicted by even a minimally trained model, significantly improved their accuracy and speed on a labeling task. However, the productivity benefits of labeling assistance are limited because the labeler still needs to consider and manually label each example.

A natural extension of assisted labeling is *semi-automated labeling*. In a semi-automated labeling paradigm the predictive model (the *machine labeler*) not only assists the human when deciding on which labels to apply, but is also capable of automating some portion of the labeling task itself. During assisted labeling, the human labeler implicitly inspects and corrects the predictions of the machine labeler, which in turn allows the machine labeler to learn and improve. In the semi-automatic paradigm this process continues until the human is satisfied with the labeling performance of the machine labeler and delegates labeling of the remaining data without further intervention.

Our semi-automated labeling approach relies on several foundational technologies. Active Learning (AL) Settles (2009) is a well established framework for minimizing labeling effort. The idea behind AL is to prioritize the most informative or "valuable" data to be considered for human labeling. Value is often determined by model uncertainty, that is, if a model (trained on existing labeled data) is uncertain about an unlabeled instance, it is likely valuable as a training example. AL is typically executed iteratively to keep the model, and thus the uncertainty estimations up to date with labeling activity. In our semi-automated labeling system, AL provides the mechanism which focuses human labeling attention on the most informative examples, thus maximizing the learning potential of the machine labeler. The sooner the most uncertain examples are labelled, from the machine's perspective, the better its performance, and the more likely automation will be feasible.

To implement a machine labeler we use a semi-supervised learning (SSL) Van Engelen and Hoos (2020) algorithm. SSL refers to a family of algorithms which specialize in learning from labeled and unlabeled data, both of which co-exist in the data labeling context. Using a semi-supervised learning algorithm allows the machine labeler to infer labels from the structure of unlabeled data, in addition to labeled examples. Combining active learning with a semi-supervised learning algorithm provides a setting in which the machine labeler can optimally learn from human labeling decisions, and importantly, human labeling effort is minimized.

In order for the human labeler to understand the performance of the machine labeler, and reason about the viability of automation, various measures of agreement are tracked and presented in a system of checkpoints. Agreement metrics describe the variance between human labeling decisions and machine predictions over the course of the labeling task. The goal is to express the performance of the machine labeler. At each checkpoint, the labeler is given the option to auto-label the remainder of the data, which corresponds to delegating labeling to the machine labeler.

2. Related Work

2.1. Algorithms & Frameworks

Significant effort has been devoted to the development of algorithms and frameworks to reduce human labeling effort and improve label quality.

As previously mentioned, Active Learning is a well established approach for minimizing human labeling effort. A selection heuristic is used to identify only the most informative unlabeled examples to present to a human for labeling. Popular selection heuristics include uncertainty, expected error reduction Roy and McCallum (2001) and query by committee Freund et al. (1997). More advanced approaches use reinforcement learning Fang et al. (2017) and deep learning Liu et al. (2018) to train custom selection models. Active Learning is particularly relevant to assisted and semi-automated labeling as it provides a framework to optimize learning, and to order the labeling task by difficulty.

Semi-supervised learning is an extension of supervised learning where algorithms simultaneously learn from both labeled and unlabeled data. A popular approach is pseudo-labeling Lee et al. (2013) which involves using the predictions of a supervised model to create more labeled data by treating predictions themselves as labels, known as pseudo-labels. The model is retrained using the original labeled data and pseudo-labeled data.

Refinements of the basic algorithm include the use of confidence thresholds and soft labels Arazo et al. (2020). Transductive semi-supervised learning describes a sub-family of graph based algorithms based on propagation of label signals within an affinity graph constructed from both labeled and unlabeled data. Label propagation Zhu and Ghahramani (2002) and label spreading Zhou et al. (2004) are two popular implementations. Label spreading being more robust to label noise via soft clamping of labeled nodes in the graph during convergence. Semi-supervised learning is particularly relevant to labeling assistance due to the co-existence of labeled and unlabeled examples, and graph based approaches are appropriate due to the capability of handling large amounts of unlabeled data.

Zhang et al. (2014) introduced the notion of cooperative learning which combines active learning and semi supervised learning. The idea of cooperative learning is to share the labeling work between human labelers and a machine so that examples predicted with insufficient confidence are subject to human labeling, and those with high confidence values are automatically labeled. Baur et al. (2020), in the context of social signal annotation, reported a reduction of manual labeling to 5/8 of data, corresponding to a saving of 2.5h of human effort, when using a cooperative learning system. Our work is complimentary to cooperative learning. However, we focus on the use of labeling difficulty and human-machine agreement as reasoning tools, rather than model confidence. Our work also considers delegation of the labeling task to the machine labeler.

2.2. Interactive Machine Learning

Semi-automated labeling falls into the larger category of interactive machine learning or human-in-the-loop systems, in which human users are integrated into the machine learning system Amershi et al. (2014). More specifically, our semi-automated labeling approach involves *interactive labeling* or *interactive annotation* Knaeble et al. (2019). Interactive annotation using active learning has shown to perform better than manual annotation Schreiner et al. (2007); Benato et al. (2021).

From a feature perspective, Sun et al. (2017) studied visualizations of the learning process during interactive labeling, and found that visual representations of model performance, such as model predictions and an agreement graph tracking the human labeling decision vs. a classifier's prediction, improved users' understanding and motivation, and helped to avoid redundant labeling. Rosenthal and Dey (2010) found that model predictions and uncertainty scores helped to improve labeling accuracy.

In a study of human-machine decision making, Lai and Tan (2019) demonstrated that providing humans with machine predictions significantly improved human decision-making performance in a deception-detection task. They found that showing predictions of a model resulted in a 21% accuracy improvement and showing the predictions along with confidence scores resulted in a 46% relative improvement. In a similar study, Zhang et al. (2020) measured improvement in users' trust when provided with prediction and confidence scores but found no significant improvement in accuracy. The authors attributed the insignificant accuracy gain to the human and AI having little performance divergence on the task. Finally Desmond et al. (2021) studied the effect of AI assistance on data labeling performance, discovering significant improvements in accuracy and labeling speed when labelers were assisted by a predictive model.

2.3. Commercial Tools

In recent years the idea of using a predictive model to either assist or automate labeling is becoming more prevalent, even in commercial offerings. Examples of such systems are Clarifai[1], Labellerr[2], Amazon Sage Maker Ground Truth[3] and IBM Cloud Annotations[4]. Our work differs in that we treat the labeling process as an ongoing collaboration between human and machine, rather than a discrete function applied at some fixed point in the labeling task.

3. Semi-Automated Data Labeling

Semi-automated data labeling frames the labeling problem as a collaboration between a human labeler and a machine labeler (implemented as a predictive model). Figure 1 presents a high level overview of the semi-automatic labeling flow. At the core of the approach is a human-in-the-loop process driven by a semi-supervised predictive model (Label Spreading) and an active learning selector (Min-margin). The active learning selector prioritizes the most informative (uncertain) examples to present to the labeler at each iteration, as defined by the predictions of the model. This strategy provides a predictable difficulty gradient throughout the labeling task. The agreement between Human decisions and machine predictions are tracked and presented via a system of checkpoints which keep the labeler aware of how the machine is performing on the labeling task. When the labeler is satisfied with the performance of the machine, they can proceed to delegate the remainder of the labeling task to the machine (Auto Labeling).

In the following sections, we describe our semi-automated labeling system alongside experimental evaluation results, and interface design discussion. Experimental evaluation was performed on a set of four datasets. Details are provided in Table 1. Text datasets were encoded using the Universal Sentence Encoder Cer et al. (2018).

Dataset	Examples	Labels	Description
stack-exchange-5k	5000	15	A subset of the stack exchange dataset [5]
chatbot	2159	107	A customer assistance chat bot dataset
spam	5572	2	SMS Spam Collection Data Set[6]
mnist-5k	5000	10	A subset of the MNIST dataset

Table 1: Details of evaluation datasets.

1. https://www.clarifai.com/
2. https://www.labellerr.com/
3. https://aws.amazon.com/sagemaker/groundtruth/
4. https://cloud.annotations.ai/
5. The stack exchange dataset is attributed to the Stack Exchange Network. https://stackexchange.com
6. https://archive.ics.uci.edu/ml/datasets/sms+spam+collection#

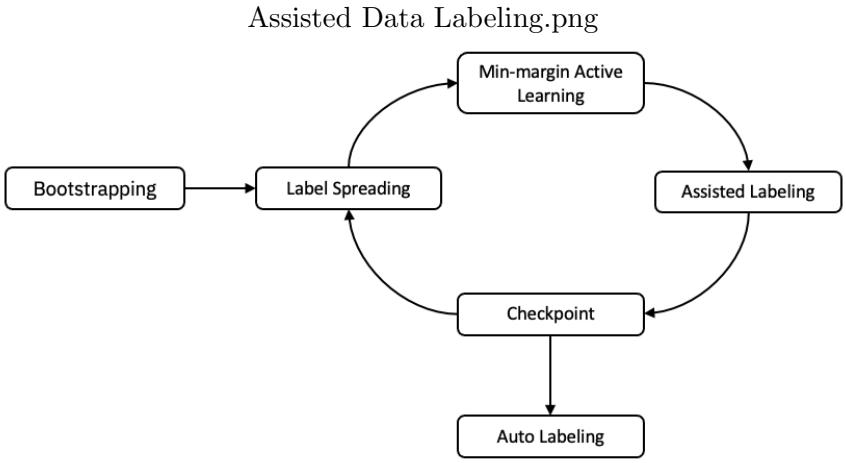

Figure 1: An abstract representation of the semi-automated data labeling flow.

3.1. Bootstrapping

At the core of semi-automated labeling process is a model that predicts labels, but this model requires an initial set of labeled data to begin making the predictions. In some cases the labeler may have access to an initial set of labeled data, but in the absence of existing labeled data, the system needs to be bootstrapped.

The role of bootstrapping is to find a minimal set of representative examples from the unlabeled data for an initial round of labeling. The set should be minimal to reduce the human effort involved, but representative to express the nature of the labeling task and thus maximize the quality of the label predictions.

Relying on the cluster assumption Chapelle and Zien (2005) our bootstrapping algorithm partitions the data using K-means clustering. K is provided by the labeler and is interpreted as an approximation of the number of labels appropriate for the labeling problem. K may be increased to improve the likelihood that all latent classes will be discovered. Representative examples are selected based on closest proximity to the K cluster centers discovered in the cluster stage. Typically, a single example closest to each cluster center is sufficient for label spreading to work well. Table 2 presents a comparison of the K-means bootstrapping algorithm versus a random approach (where a set of examples are randomly selected). A label spreading model trained using the representative examples selected by the K-means algorithm provide better predictive performance than a random selection, consistent across all datasets.

From an interface point of view, boostrap examples are presented to the labeler without assistance from the machine labeler. Figure 2 shows the "unassisted" labeling interface used during bootstrapping.

3.2. Iterative Refinement

After bootstrapping, the remainder of the labeling process occurs as a human-machine labeling loop, where the machine labeler learns from the labeling decisions made by the human labeler. At the beginning of each iteration, the machine labeler predicts label distributions

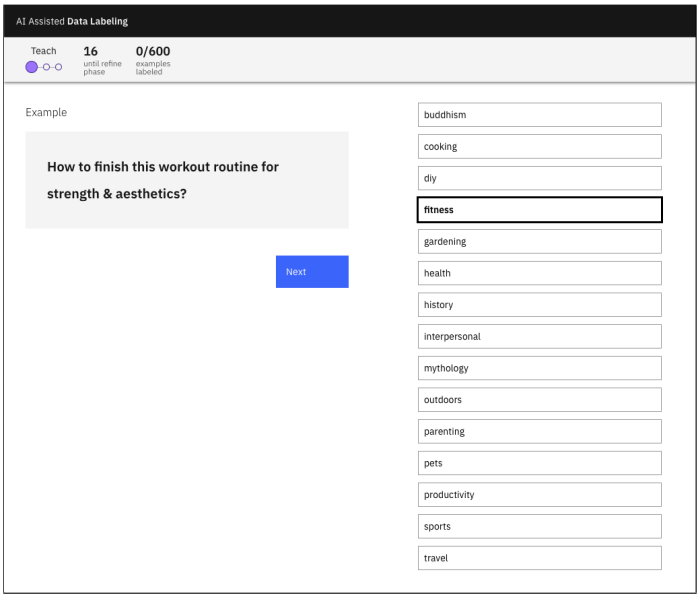

Figure 2: The unassisted labeling interface used for bootstrap labeling. The example is presented to the left of the screen, and the set of labels are presented in an alphabetical list to the right. The interface also provides progress information.

Dataset	Sample size (K)	Random	K-means center
stack-exchange-5k	15	0.293	**0.558**
chatbot	107	0.28	**0.415**
spam	10	0.892	**0.934**
mnist-5k	10	0.416	**0.582**

Table 2: Prediction accuracy of the label spreading algorithm on remaining unlabeled data after initial bootstrapping using random selection vs. the K-means center algorithm. The bootstrap parameter K is set to the number of labels in the original dataset (except for the spam dataset where it is set to 10). Results are averaged over 3 random shuffles of the data.

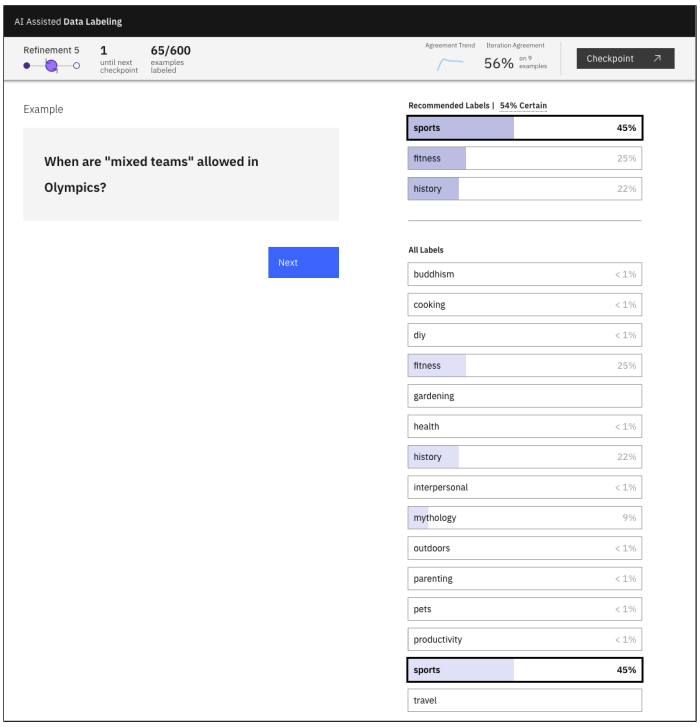

Figure 3: The assisted data labeling interface used during iterative refinement.

for all remaining unlabeled data. The active learning heuristic is then applied, selecting a batch of the most uncertain predictions. The batch of unlabeled data, along with the corresponding predicted labels, are presented to the human labeler for "assisted" labeling. Showing predictions helps the human labeler to decide on the correct label by narrowing their focus to the most probable subset.

Figure 3 shows the assisted labeling interface that is used during iterative refinement. Label predictions are displayed in a dedicated "Recommended Labels" list, limited to the top 3 predictions. This allows the labeler to very quickly understand and access the top predictions of the machine sorted by confidence (prediction probability). In addition, an alphabetical list of all labels is provided, annotated with the predicted label distribution (purple bars overlaid on the list of labels). The full label display is provided so that the labeler can fall back on spatial memory to find labels, in the event that the desired label is not recommended by the machine labeler. The top predicted label is always pre-selected in the interface, so that as the performance of machine labeler improves, the mechanical task of labeling becomes more fluent for the human labeler. We consider this a form of automation within the labeling interface itself, and an intermediate step towards automation.

3.2.1. Label Spreading

Label predictions are computed using label spreading. As mentioned in 2.1 label spreading is a transductive graph based semi-supervised learning algorithm. The algorithm works by propagating label signals through an affinity graph derived from both the labeled and

unlabeled data. This is essentially a graph weighted by similarity. At each step in the algorithm convergence process the graph laplacian matrix is used to propagate label signals. A system of soft clamping is applied to retain ground truth on labeled examples, which may otherwise be lost as the algorithm converges.

The output of the label spreading algorithm[7] is a label probability distribution for each remaining unlabeled example. Label spreading is particularly well suited for the data labeling context as it works reliably even with a very small set of labeled examples, and scales to large data sets.

3.2.2. Active learning

A key aspect of the semi-automated labeling workflow is to maintaining a descending gradient of labeling difficulty. Having the human labeler focus on the most difficult examples upfront, improves the performance of the machine labeler as soon as possible. This is fundamental principle of uncertainty based active learning. The gradient also helps the labeler to reason about delegation of the remaining labeling task. If the remaining labeling task is relatively easier, then the observed performance on the completed labeling tasks provides an approximate lower bound expectation of what to expect upon delegation. Put another way, if the label prediction model is performing well on earlier iterations, it is likely to do as well or better in subsequent iterations and the human labeler can feel confident in delegating the remainder of examples.

A min-margin active learning selector is used at each iteration of the labeling loop to select a batch of examples from the remaining unlabeled pool. The min-margin heuristic prioritizes examples with the smallest probability margin between the top two predicted labels. Figure 4 demonstrates the performance (accuracy on the remaining unlabeled data) of a min-margin selector versus a random selector across 150 simulated labeling iterations. In all scenarios the min_margin selector achieves higher accuracy at predicting the correct labels on the remaining unlabeled data in each iteration, when compared to a random selector.

3.3. Checkpoints

An important question in a semi-automated labeling system is when to automate the labeling task. After each refinement iteration (the size of which is configurable per labeling problem) the labeler is presented with a checkpoint (Figure 5). The checkpoint is designed to allow the human labeler to consider the current performance of the machine labeler and decide if the remainder of the unlabeled data should be automatically labeled (delegated), or if further refinement is necessary. Research has shown that in order for humans to trust automation, they need high 'temporal specificity', which means that they need to be updated frequently on the performance of the system Lee and See (2004). By ensuring that the labeler is regularly updated on the performance of the model, the checkpoints provide 'temporal specificity' and manage user expectations throughout the interaction, a guiding design principle for interactive machine learning Dudley and Kristensson (2018).

Checkpoints also address improved user engagement by breaking up the otherwise monotonous task of data labeling. Research on crowd work has shown that 'micro-diversions'

7. https://scikit-learn.org/stable/modules/generated/sklearn.semi_supervised.LabelSpreading.html

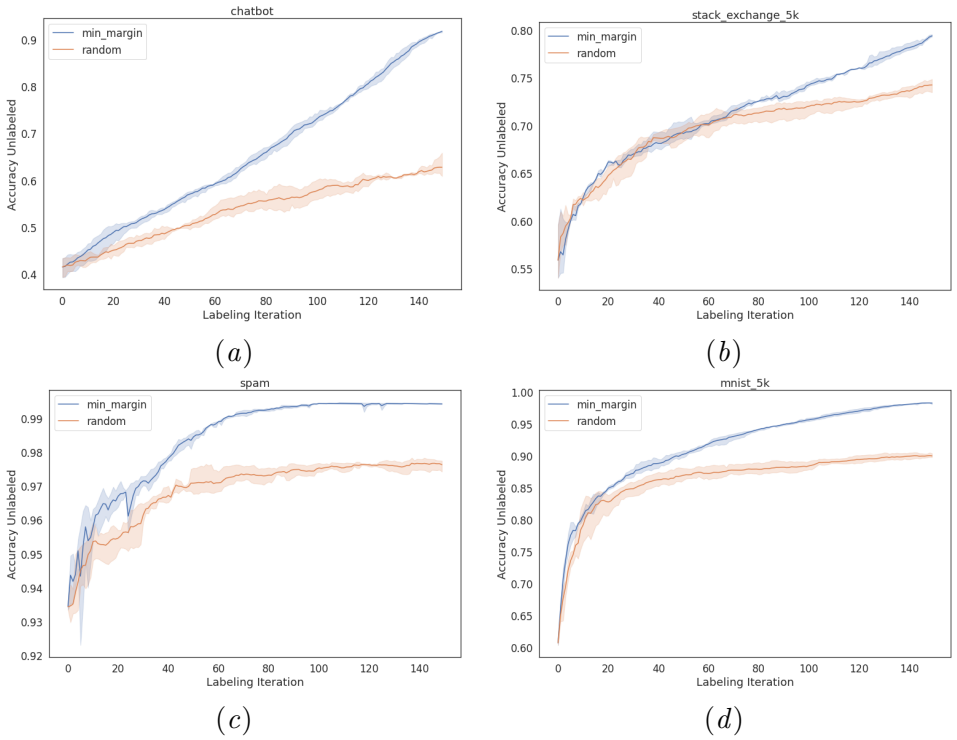

Figure 4: Comparing the performance of a min_margin active learning selector (blue) vs. a random selector (orange) over the course of 150 labeling iterations. Labeling simulated using an oracle. (a) chatbot dataset, (b) stack_exchange_5k dataset, (c) spam dataset, (d) mnist_5k dataset.

have been helpful in user engagement in other types of crowd work Dai et al. (2015); Rzeszotarski et al. (2013).

3.3.1. Metrics

At each checkpoint three metrics are tracked via the "Agreement Trend" graph shown in Figure 5. The metrics are calculated and presented per iteration of the labeling loop prior to the checkpoint. The chart plots the human-machine labeling agreement, human-machine agreement similarity, and labeling task difficulty. Each of the metrics play a role in the delegation decision.

Agreement Agreement simply counts the number of times the human and machine labeler agree on the correct label, aggregated per refinement iteration. The top predicted label is taken as the machine labelers choice. The agreement metric helps the human labeler to understand the performance of the machine labeler relative to their own labeling decisions. In the checkpoint in Figure 5, agreement between human and machine labeler increases over the span of 14 iterations, and is consistently 100% over the last three refinement iterations. This would be a scenario where delegation may be an option.

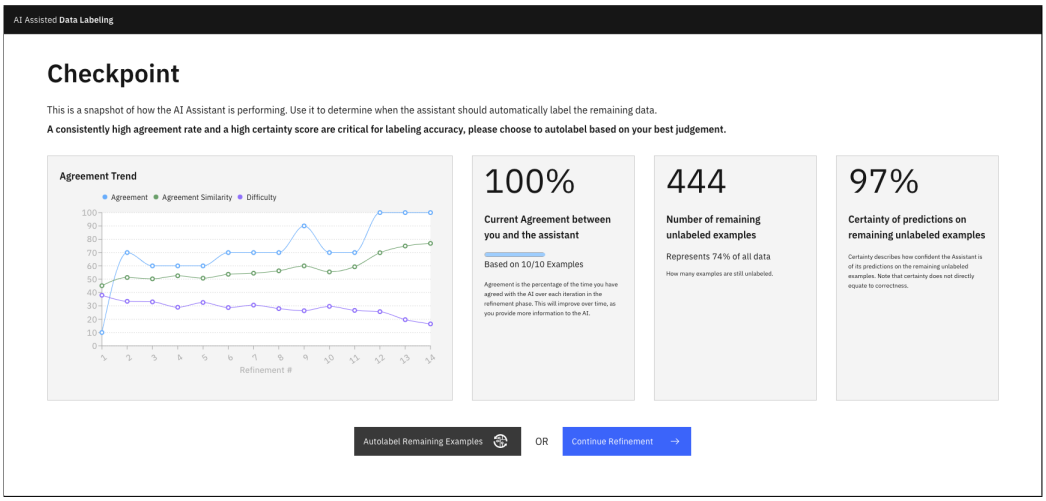

Figure 5: A checkpoint displaying the state of the labeling task. In this example the size of each labeling refinement iteration is 10 examples. The labeler has the option to delegate the remainder of the labeling task (444 examples) or continue to the next refinement step.

Agreement similarity Agreement similarity is also intended to indicate agreement. However, unlike the simple agreement metric which only considers the top predicted label, agreement similarity considers the overall distribution of the machine prediction. The metric is calculated as the inverse Jensen-Shannon distance [8] between the predicted label distribution of the machine labeler and the human selected label, one hot encoded as a distribution. Intuitively the agreement similarity provides a smoother representation of agreement, regardless of the actual correctness of the machine labelers prediction. In the checkpoint in Figure 5, the agreement similarity is trending upwards suggesting increasing performance of the machine labeler.

Labeling Difficulty The "difficulty" of each labeling iteration is calculated as the mean scaled entropy [9] of the predicted label distributions for all examples in the iteration. This metric is expected to trend downwards as the machine labeler learns and improves at the task. This downward trend is shown in the checkpoint in Figure 5. Tracking and presenting difficulty is important to demonstrate to the human labeler that future labeling work tends to be easier and more predictable due to the uncertainty based data prioritization in use.

4. User Evaluation

Our semi-automated labeling system was made available as an online web application for audience use at the NeurIPS 2020 demonstration track. Over the course of the conference and the subsequent month a total of 191 people used the system. Of the 191 users, 52

8. https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.jensenshannon.html
9. Scaled entropy refers to dividing calculated entropy by log k, so that results then fall within the unit interval 0-1

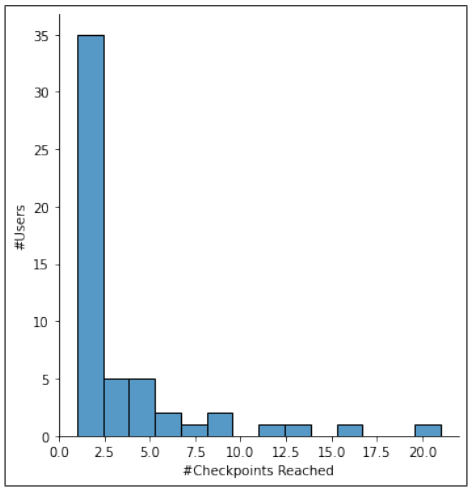

Figure 6: Number of users who reached each checkpoint

(27.2%) completed the full semi-automated labeling experience. This involved labeling a total set of 600 short text examples, using 21 labels, with the help of a machine labeler.

We wanted to understand when users chose to auto-label the remaining unlabeled data (delegate to the machine) and found that users who completed the experience on average labeled 48 of the 600 data points manually ($SD = 41$). Most participants chose to auto-label after manually reviewing 10-30 items (1-3 checkpoints), as shown in Figure 6.

A short survey at the completion of the experience focused on understanding users' choices about when to auto-label. Of the 52 users who finished the experience, 21 filled out our short survey at the conclusion of the labeling tasks. The most popular way users chose when to auto-label was based on their agreement with the model from the checkpoint (11/21 users). Several participants used the suggested labels for examples during labeling and the model's certainty presented in the checkpoint (2/21 users and 3/21 users, respectively). One participant wrote that "agreement rate stopped increasing," indicating that they were keeping track of the changes in performance over time.

The data collected from the conference demo has several limitations. Conference attendees are not necessarily motivated to perform labeling to the best of their abilities. As expected for usage during a conference demo, 4 of the 21 participants reported just wanting to be done labeling. One of these participants noted that "for production use would have reached for higher confidence." Furthermore, we chose to use a subset of the stack-exchange dataset for the demo in part because the model confidence and accuracy increased with a reasonable amount of manual labeling. Different datasets and tasks would impact the users' experience, as the speed of improvement of the machine labeler would differ. Regardless, the demo participants' data does provide early information on how people might use this type of system, in particular when people might choose to allow a system to complete the remaining data labeling tasks for a dataset.

5. Conclusion

In this paper we have described a semi-automated data labeling workflow and tool. We discussed the algorithmic components of the system, and the corresponding user experience. We also presented some experimental validation of the work, and shared some initial user evaluation based on a live demo deployed at the NeurIPS 2020 conference. Our work builds upon interactive machine learning techniques and presents a novel structure for semi-automatic data labeling. The system is designed to keep users informed about the machine labelers progress and enable them to defer the remaining labeling to the system, without inspecting the data or assigned labels. We expect that this system structure will encourage humans to label the more challenging items and offload the easy items entirely to the machine, reducing overall labeling effort.

Going forward our plan is to further evaluate semi-automated labeling as a working paradigm, focusing on measuring the utility of the approach across various datasets and tasks, and understanding how users perceive and interact with such a system. We also believe that cooperative human-machine labeling may have particular value in collaborative settings, involving a team of labelers. The assistance of a machine labeler in such a setting, may help the team to converge on a consistent interpretation of the labeling space, increasing inter rater reliability and overall label quality.

References

Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4):105–120, 2014.

Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020.

Tobias Baur, Alexander Heimerl, Florian Lingenfelser, Johannes Wagner, Michel F Valstar, Björn Schuller, and Elisabeth André. Explainable cooperative machine learning with nova. *KI-Künstliche Intelligenz*, pages 1–22, 2020.

Bárbara C Benato, Jancarlo F Gomes, Alexandru C Telea, and Alexandre X Falcão. Semi-automatic data annotation guided by feature space projection. *Pattern Recognition*, 109, 2021. ISSN 00313203.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

Olivier Chapelle and Alexander Zien. Semi-supervised classification by low density separation. *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, 57-64 (2005)*, 01 2005.

Peng Dai, Jeffrey M Rzeszotarski, Praveen Paritosh, and Ed H Chi. And now for something completely different: Improving crowdsourcing workflows with micro-diversions. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 628–638, 2015.

Michael Desmond, Zahra Ashktorab, Michelle Brachman, Kristina Brimijoin, Evelyn Duesterwald, Casey Dugan, Catherine Finegan-Dollak, Michael Muller, Narendra Nath Joshi, Qian Pan, and Aabhas Sharma. Increasing the speed and accuracy of data labeling through an ai assisted interface. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, 2021.

John J. Dudley and Per Ola Kristensson. A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems*, 8, 2018.

Meng Fang, Yuan Li, and Trevor Cohn. Learning how to active learn: A deep reinforcement learning approach. *arXiv preprint arXiv:1708.02383*, 2017.

Yoav Freund, H Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Machine learning*, 28(2):133–168, 1997.

Merlin Knaeble, Mario Nadj, and Alexander Maedche. Oracle or teacher? a systematic overview of research on interactive labeling for machine learning. In *Internationaler Kongress Für Wirtschaftsinformatik 2020*, 2019.

Vivian Lai and Chenhao Tan. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 29–38, 2019.

Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 2013.

John D Lee and Katrina A See. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80, 2004.

Ming Liu, Wray Buntine, and Gholamreza Haffari. Learning how to actively learn: A deep imitation learning approach. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1874–1883, 2018.

Stephanie L Rosenthal and Anind K Dey. Towards maximizing the accuracy of human-labeled sensor data. In *Proceedings of the 15th International Conference on Intelligent User Interfaces*, pages 259–268, 2010.

Nicholas Roy and Andrew McCallum. Toward optimal active learning through monte carlo estimation of error reduction. *ICML, Williamstown*, pages 441–448, 2001.

Jeffrey Rzeszotarski, Ed Chi, Praveen Paritosh, and Peng Dai. Inserting micro-breaks into crowdsourcing workflows. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 1, 2013.

Chris Schreiner, Harry Zhang, Claudia Guerrero, Kari Torkkola, and Keshu Zhang. A Semi-Automatic Data Annotation Tool for Driving Simulator Data Reduction. *Driving Simulation Conference, North America*, 2007.

Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

Yunjia Sun, Edward Lank, and Michael Terry. Label-and-learn: Visualizing the likelihood of machine learning classifier's success during data labeling. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, pages 523–534, 2017.

Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, 2020.

Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 295–305, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372852. URL https://doi.org/10.1145/3351095.3372852.

Zixing Zhang, Eduardo Coutinho, Jun Deng, and Björn Schuller. Cooperative learning and its application to emotion recognition from speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(1):115–126, 2014.

Dengyong Zhou, Olivier Bousquet, Thomas N Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in neural information processing systems*, pages 321–328, 2004.

Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical report, 2002.