Enhancing Human-in-the-Loop Learning for Binary Sentiment Word Classification

Belen Martin-Urcelay[1], Christopher J. Rozell[1], Matthieu R. Bloch[1]

***Abstract*— While humans intuitively excel at classifying words according to their connotation, transcribing this innate skill into algorithms remains challenging. We present a human-guided methodology to learn binary word sentiment classifiers from fewer interactions with humans. We introduce a human perception model that relates the perceived sentiment of a word to the distance between the word and the unknown classifier. Our model informs the design of queries that capture more nuanced information than traditional queries solely requesting labels. Together with active learning strategies, our approach reduces human effort without sacrificing learning fidelity. We validate our method through experiments with human data, demonstrating improved accuracy in binary sentiment word classification.**

I. INTRODUCTION

While humans are adept at classification, articulating the underlying rules that lead to these classifications remains elusive [1], leading to a gap between implicit human knowledge and explicit classification rules of learning algorithms. This gap makes it challenging to integrate human expertise into learning algorithms and to develop machines capable of making decisions informed by human experts. Addressing this challenge is critical in a wide range of applications, including medical diagnosis [2], [3], robotics [4], [5], and word sense disambiguation [6].

We focus here on the binary sentiment word classification task [7]. While humans can instinctively label words according to their connotation as positive (e.g., healthy) or negative (e.g., scary) [8], articulating the defining attributes of each category proves challenging for most people. The challenge is compounded by the need to transfer this implicit human knowledge in a form that is understandable to both humans and algorithms. As a result, the role of humans is often limited to serving as oracles that label data points [9], [10], [11].

Reducing the number of queries is crucial not just for enhancing computational efficiency but also for minimizing human annotator fatigue, thereby improving data quality. Current strategies in sentiment classification [12], [13] incorporate active learning [14], [15] to select the most informative queries for human annotation. However, since humans only have the limited role of providing labels, the amount of information obtained from each query is low [16]. In the case of a binary classifier, the maximum information gain per query is only 1 bit. Therefore, a high number of interactions and substantial labeling effort are required.

Recognizing the inherent inefficiency, we propose richer query forms that retain comprehensibility for both humans and algorithms, while increasing the information gain. We extend active learning frameworks to adapt to these enhanced query types, further reducing the frequency of human-computer interactions required.

The main contributions are as follows:

- We introduce a human response model that exploits a newly observed linear relationship between the perceived sentiment associated to a word and the distance of the word to the sentiment classification boundary. This model enhances the understanding and prediction of human word selection behaviors.
- We design more informative queries for sentiment classification by combining label requests with word selection. We also propose a computationally efficient strategy to update classifier beliefs based on these enriched responses and to actively select informative queries.
- We empirically validate our approach through experiments involving human data. We demonstrate the effectiveness of our method in enhancing sentiment classification performance.

II. MODEL

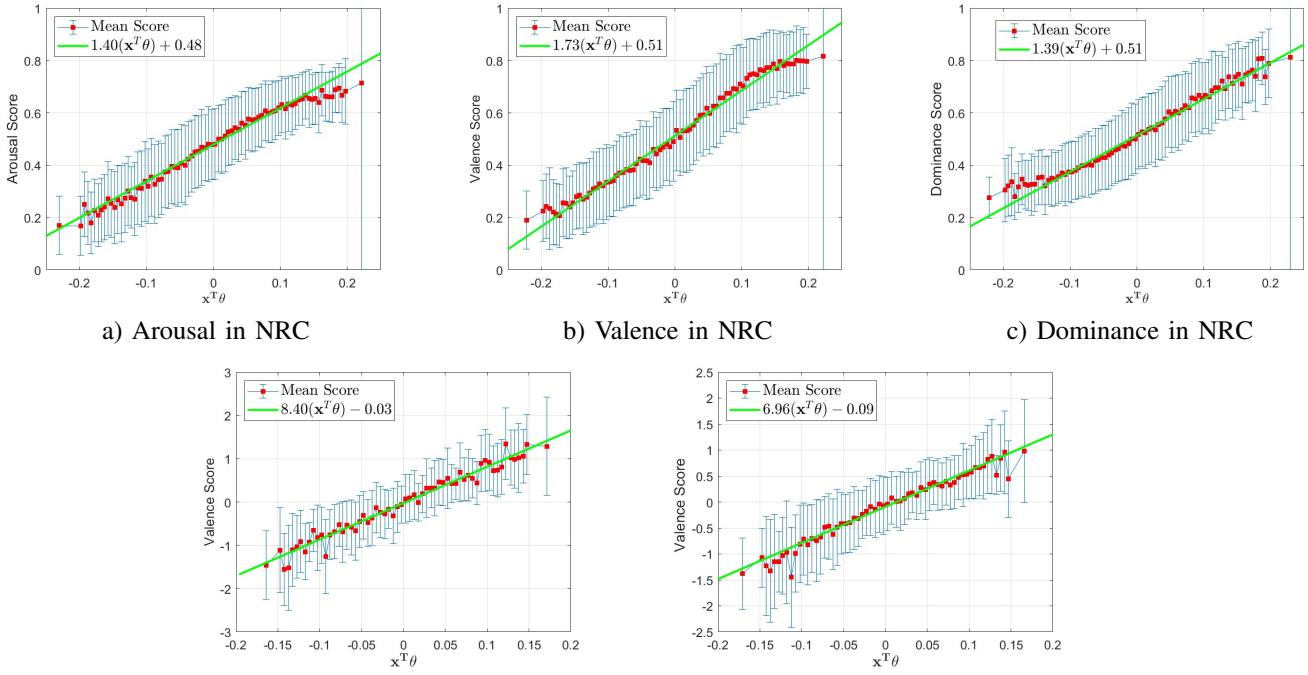
Our primary goal is to learn a classifier separating words according to their connotation using the minimum number of interactions with humans. Specifically, we want to learn a linear binary classifier $\boldsymbol{\theta} \in \mathbb{R}^d$ that correctly labels words $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$ as positive or negative according to $y = \text{sign}(\boldsymbol{\theta}^T \mathbf{x})$. An established method for learning this implicit knowledge, is to use humans as oracles within the framework of the Bradley-Torrey model [19]. This model encapsulates the probability of a word being labeled as positive ($y = 1$) in the form of a logistic function

$$\mathbb{P}\left[y = 1 | \mathbf{x}\right] = \left(1 + \exp\left(w(\boldsymbol{\theta}^T \mathbf{x})\right)\right)^{-1}, \tag{1}$$

with $w \in \mathbb{R}$ representing the inverse of the scale parameter.

To obtain more information from the humans, we propose not only soliciting labels but also asking participants to choose a word from a pool according to a query. This requires a model that connects word embeddings with the classifier in a manner both intuitive for humans and quantifiable for computational analysis. To force this relationships, previous works [20] tailor embedding spaces to specific

[1]Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, Georgia, U.S.A. burcelay3@gatech.edu, crozell@gatech.edu, matthieu.bloch@ece.gatech.edu



a) Arousal in NRC b) Valence in NRC c) Dominance in NRC

d) Valence of adjectives in SocialSent e) Valence of frequent words in SocialSent

Fig. 1: Relationship between the score of words, given by the National Research Council Canada (NRC) [17] or SocialSent [18] lexicons, and the distance of its embedding to the ground truth classifier. We observe there is an approximately linear relationship.

tasks. Unlike these approaches, our model seeks to uncover relationships that hold on pre-existing word embeddings in a general way, for a wide range of datasets and classification tasks.

Intuitively, we expect humans to be dubious when asked to label word instances close to the boundary between classes, and to find words more positive or negative as we move farther from the hyperplane in opposite directions. Heuristically, the intuition holds true for several lexicons in a preexisting word embedding [21]. Previous work [22] shows that there are three fundamental dimensions of meaning: *valence*, representing the spectrum between positivity and negativity; *arousal*, representing the contrast between active and passive emotions; and *dominance*, capturing the power dynamics from submissive to dominant. In all three dimensions, Figure 1 shows how the scores of words, as provided by the NRC dataset [17], vary with the distance to the Minimum Mean-Square Error (MMSE) classifier. More importantly, we observe an approximately linear relationship between the score of a word and the inner product of its embedding with the classifier. As shown in Figure 1d and Figure 1e, this behavior is consistent across distinct and independently gathered datasets [18] of valence scores. These observations suggest an inherent connection between word sentiment scores and the distance of their embeddings to the classifier. To formalize this observation, we introduce Assumption II.1.

Assumption II.1. The implicit sentiment score associated with a word $\mathbf{x}$ by a human is given by

$$\text{score}(\mathbf{x}) = a\mathbf{x}^T\boldsymbol{\theta} + b + \epsilon, \tag{2}$$

where $\epsilon \sim \text{Gumbel}(\mu, \sigma)$ is a random variable distributed according to a Gumbel distribution. The scalars $a \in \mathbb{R}$ and $b \in \mathbb{R}$, which describe the linear relationship, are dataset dependent.

Consider the question q_{pos} = "Select and label the word you find most positive;" in that case, we expect humans to select the word with the highest perceived sentiment score. We model the probability that the word $\mathbf{x}_i$ is selected from a word set $\mathcal{S} = \{\mathbf{x}_j\}_{j=1}^{|\mathcal{S}|}$ as

$$\begin{aligned}\mathbb{P}[\mathbf{x}_i|\mathcal{S}, \boldsymbol{\theta}, q_{\text{pos}}] &= \mathbb{P}[\text{score}(\mathbf{x}_i) > \text{score}(\mathbf{x}_j), \forall j \neq i] \\ &= \mathbb{P}[\epsilon_j - \epsilon_i < a(\mathbf{x}_i - \mathbf{x}_j)^T\boldsymbol{\theta}, \forall j \neq i] \\ &= \frac{\exp\left(\frac{a}{\sigma}\mathbf{x}_i^T\boldsymbol{\theta}\right)}{\sum_{\mathbf{x}\in\mathcal{S}}\exp\left(\frac{a}{\sigma}\mathbf{x}^T\boldsymbol{\theta}\right)},\end{aligned} \tag{3}$$

because this is a logit choice probability [23]. Analogously, for the question q_{neg} ="Select and label the word you find most negative,"

$$\mathbb{P}[\mathbf{x}_i|\mathcal{S}, \boldsymbol{\theta}, q_{\text{neg}}] = \frac{\exp\left(-\frac{a}{\sigma}\mathbf{x}_i^T\boldsymbol{\theta}\right)}{\sum_{\mathbf{x}\in\mathcal{S}}\exp\left(-\frac{a}{\sigma}\mathbf{x}^T\boldsymbol{\theta}\right)}. \tag{4}$$

Compared to the traditional approach of only requesting the label, we gather more information from each query using questions q_{pos} and q_{neg}.

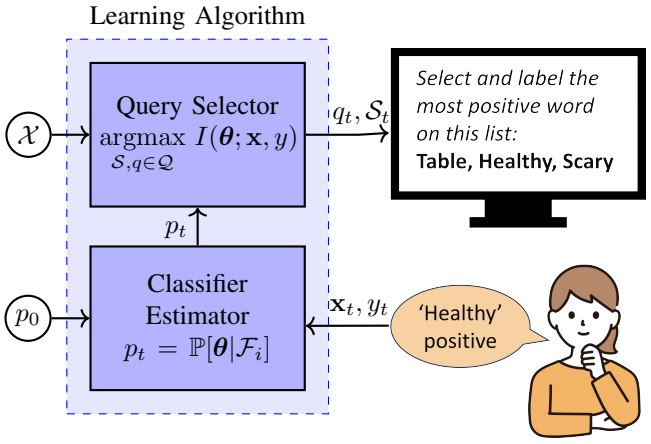


Fig. 2: Block diagram for human-in-the-loop learning for sentiment word classification. At each interaction, the human receives a query asking to select a word from a list, and provide its label. The answer to the query is used to update the estimator of the classifier and select the next query.

III. ALGORITHM

We propose an online machine learning algorithm to learn a binary classifier of words from human feedback. Figure 2 illustrates the principle of our approach. At each iteration, the learning algorithm selects the query (question $q \in \mathcal{Q} = \{q_{\text{pos}}, q_{\text{neg}}\}$ and word set $\mathcal{S}$) such that the expected human response is as informative as possible, i.e., $\{q, \mathcal{S}\}$ maximizes the mutual information [24] between the underlying classifier and the human response. This task is managed by the query selector. Next, the human answers the query by selecting a word $\mathbf{x}$ from $\mathcal{S}$ and providing its label y. The classifier estimator collects the response from the human and leverages this information to update the estimator of the classifier $\boldsymbol{\theta}$. The posterior $\mathbb{P}[\boldsymbol{\theta}|\mathcal{F}_t]$, where $\mathcal{F}_t = \{\mathbf{x}_i, y_i, q_i, \mathcal{S}_i\}_{i=1}^t$ denotes the history, is updated using Bayes rule. Namely, the classifier estimator leverages the likelihood functions in (1), (3) and (4). The approach is summarized in Algorithm 1.

Algorithm 1 Ideal Human-in-the-Loop Learning

1: **Input:** $\mathcal{X}, \mathcal{Q}, |\mathcal{S}|, \mathbb{P}[\boldsymbol{\theta}|\mathcal{F}_0], \epsilon$
2: $t = 0$
3: **while** uncertainty$(\boldsymbol{\theta}|\mathcal{F}_t) \geq \epsilon$ **do**
4: $\quad q_t, \mathcal{S}_t \leftarrow \underset{q\in\mathcal{Q}, \mathcal{S}\in\{\mathcal{X}\}^{|\mathcal{S}|}}{\text{argmax}} \mathbb{E}_{\mathbf{x},y|q,\mathcal{S}} \left[I(\boldsymbol{\theta}; \mathbf{x}, y|q, \mathcal{S}, \mathcal{F}_{t-1}) \right]$
5: $\quad \mathbf{x}_t, y_t \leftarrow$ human response to the query
6: $\quad \mathbb{P}[\boldsymbol{\theta}|\mathcal{F}_t] \propto \mathbb{P}\left[\mathbf{x}_t|\boldsymbol{\theta}, q_t, \mathcal{S}_t\right] \mathbb{P}\left[y_t|\mathbf{x}_t, \boldsymbol{\theta}\right] \mathbb{P}[\boldsymbol{\theta}|\mathcal{F}_{t-1}]$
7: $\quad t = t + 1$
8: **end while**

Since Algorithm 1 is intractable in high dimensional settings, we provide approximations to make the computations feasible in the next subsections. The tractable implementation is described in Algorithm 2.

Algorithm 2 Approximate Human-in-the-Loop Learning

1: **Input:** $\mathcal{X}, \mathcal{Q}, |\mathcal{S}|, \boldsymbol{\mu}_0, \Sigma_0, \epsilon^d, N$
2: t = 1
3: **while** $|\boldsymbol{\Sigma}_t| \geq \epsilon^d$ **do**
4: $\quad q_t \leftarrow$ sample uniformly from $\{q_{\text{pos}}, q_{\text{neg}}\}$
5: $\quad S_t \leftarrow \{\}$
6: $\quad$ **for** $i = 1, 2, ..., |\mathcal{S}|$ **do**
7: $\quad\quad \{\widehat{\boldsymbol{\theta}}_n\}_{n=1}^N \leftarrow$ sample i.i.d. from $\mathcal{N}(\boldsymbol{\mu}_{t-1}, \Sigma_{t-1})$
8: $\quad\quad \mathbf{s} \leftarrow$ select word from dictionary with Eq. (7)
9: $\quad\quad \mathcal{S}_t \leftarrow \{\mathcal{S}_t, \mathbf{s}\}$
10: $\quad$ **end for**
11: $\quad \mathbf{x}_t, y_t \leftarrow$ human response to the query
12: $\quad \boldsymbol{\mu}_t, \Sigma_t \leftarrow \boldsymbol{\mu}_{t-1}, \Sigma_{t-1}$
13: $\quad$ **while** $\boldsymbol{\mu}_t, \Sigma_t$ not converged **do**
14: $\quad\quad$ **while** $\boldsymbol{\mu}_t, \Sigma_t$ not converged **do**
15: $\quad\quad\quad \xi^2 = w^2 \mathbf{x}^T \boldsymbol{\Sigma}_t \mathbf{x} + w^2 (\mathbf{x}^T \boldsymbol{\mu}_t)^2$
16: $\quad\quad\quad \boldsymbol{\Sigma}_t^{-1} = \boldsymbol{\Sigma}_{t-1}^{-1} + 2\frac{\tanh(\xi/2)}{4\xi} w^2 \mathbf{x}\mathbf{x}^T$
17: $\quad\quad\quad \boldsymbol{\mu}_t = \boldsymbol{\Sigma}_t \left[\boldsymbol{\Sigma}_{t-1}^{-1} \boldsymbol{\mu}_{t-1} + \left(y - \frac{1}{2}\right) w\mathbf{x}\right]$
18: $\quad\quad$ **end while**
19: $\quad\quad \boldsymbol{\mu}_t, \Sigma_t \leftarrow$ update according to the selected word with (6)
20: $\quad$ **end while**
21: $\quad t = t + 1$
22: **end while**

Approximation of Belief Update

Line 6 of Algorithm 1 requires us to find the posterior

$$\mathbb{P}[\boldsymbol{\theta}|\mathcal{F}_t] = \frac{\mathbb{P}[\mathbf{x}_t|\boldsymbol{\theta}, \mathcal{S}_t, q_t]\mathbb{P}[y_t|\mathbf{x}_t, \boldsymbol{\theta}]}{\mathbb{P}[\mathbf{x}_t, y_t|\mathcal{S}_t, q_t, \mathcal{F}_{t-1}]} \mathbb{P}[\boldsymbol{\theta}|\mathcal{F}_{t-1}].$$

Unfortunately, the likelihood functions are not conjugate to the prior, so an analytical closed-form expression for the posterior is not available. To address this, we approximate the posterior using variational inference. While Black-Box Variational Inference (BBVI) [25] is commonly used in such situations, we derive closed-form variational updates, which we describe next, to achieve a lower variance approximation.

We approximate the classifier density function as a Gaussian distribution $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We may then compute the posterior mean and variance given a word label by an iterative process [26] described in lines 14 to 18 of Algorithm 2.

In a similar fashion, we approximate the classifier posterior given the word selected with a variational approach. We look for the variational distribution $q(\boldsymbol{\theta}) \sim \mathcal{N}(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)$ closest, in terms of the Kullback-Leibler (KL) distance, to the true posterior. This is equivalent to finding the distribution that maximizes the log Evidence Lower BOund (ELBO) [27]

$$\text{ELBO}(q) = -\,\text{KL}(q(\boldsymbol{\theta})\|p(\boldsymbol{\theta})) + \mathbb{E}_{\boldsymbol{\theta}\sim q}\left[K \mathbf{x}_s^T \boldsymbol{\theta}\right] - \mathbb{E}_{\boldsymbol{\theta}\sim q}\left[\log \sum_{j=1}^{|\mathcal{S}|} \exp\left(K \mathbf{x}_j^T \boldsymbol{\theta}\right)\right], \quad (5)$$

where $\mathbf{x}_s$ is the word selected by the human, the prior is

$p(\boldsymbol{\theta}) \sim \mathcal{N}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$ and

$$K = \begin{cases} \frac{a}{\sigma}, & q = q_{\text{pos}} \\ -\frac{a}{\sigma}, & q = q_{\text{neg}} \end{cases}.$$

The first term in (5) is the KL divergence between two Gaussian distributions,

$$\begin{aligned}\mathrm{KL}(q||p) =& \frac{1}{2}\left[\log\frac{|\boldsymbol{\Sigma_p}|}{|\boldsymbol{\Sigma_q}|} - d + \boldsymbol{\mu_q}^T\boldsymbol{\Sigma_p}^{-1}\boldsymbol{\mu_q} + \boldsymbol{\mu_p}^T\boldsymbol{\Sigma_p}^{-1}\boldsymbol{\mu_p}\right] \\ & - \boldsymbol{\mu_q}^T\boldsymbol{\Sigma_p}^{-1}\boldsymbol{\mu_p} + \frac{1}{2}tr\left\{\boldsymbol{\Sigma_p}^{-1}\boldsymbol{\Sigma_q}\right\}.\end{aligned}$$

Because of the linearity of the expectation, we compute the second term in (5) as

$$\mathbb{E}_{\boldsymbol{\theta}\sim q}\left[K\mathbf{x}_s^T\boldsymbol{\theta}\right] = K\mathbf{x}_s^T\boldsymbol{\mu}_q.$$

The third term in (5) has no closed-form solution, but following [28], we apply Jensen's inequality to obtain a lower bound

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\theta}\sim q}\left[\log\sum_{j=1}^{|\mathcal{S}|}\exp\left(K\mathbf{x}_j^T\boldsymbol{\theta}\right)\right] \\ &\qquad \geq \log\sum_{j=1}^{|\mathcal{S}|}\exp\left(K\mathbf{x}_j^T\boldsymbol{\mu}_q + 0.5\mathbf{x}_j^T\boldsymbol{\Sigma}_q\mathbf{x}_j\right).\end{aligned}$$

We approximate the posterior distribution given the word selection as a Gaussian distribution whose mean and covariance are obtained by maximizing the ELBO upperbound

$$\begin{aligned}\{\boldsymbol{\mu}, \boldsymbol{\Sigma}\} = &\operatorname*{argmin}_{\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q} \mathrm{KL}(q||p) - K\mathbf{x}_s^T\boldsymbol{\mu}_q \\ &+ \log\sum_{j=1}^{|\mathcal{S}|}\exp\left(K\mathbf{x}_j^T\boldsymbol{\mu}_q + 0.5\mathbf{x}_j^T\boldsymbol{\Sigma}_q\mathbf{x}_j\right). \quad (6)\end{aligned}$$

Lines 12 through 20 of Algorithm 2 approximate the posterior by accounting for the human response. We first update the posterior according to the label received. Then we update the posterior according to the selected word with (6). We repeat both updates until convergence.

Active Learning Heuristic for Query Selection

As line 4 of Algorithm 1 indicates, we aim to select the query that, in expectation, provides the maximum information about the true classifier. This requires us to compute the posterior over every possible combination of word set, query, label and word selection. Despite using a belief approximation, the sheer number of possible combinations renders this method computationally infeasible.

We propose an approximation based on query by committee [29]. At each iteration t, we sample N particles $\widehat{\boldsymbol{\theta}}_n \overset{\text{i.i.d}}{\sim} \mathbb{P}[\boldsymbol{\theta}|\mathcal{F}_t]$. Next, we maximize the disagreement between the prediction of each particle and the mean prediction among all particles as,

$$\begin{aligned}\{q, \mathcal{S}_t\} = \operatorname*{argmax}_{q\in\mathcal{Q}, \mathcal{S}\in\{\mathcal{X}\}^{|\mathcal{S}|}} \; & H\left(\frac{1}{N}\sum_{n=1}^{N}\left[\mathbf{x}_t, y_t\middle|\widehat{\boldsymbol{\theta}}_n, q, \mathcal{S}\right]\right) \\ & - \frac{1}{N}\sum_{n=1}^{N} H\left(\mathbf{x}_t, y_t\middle|\widehat{\boldsymbol{\theta}}_n, q, \mathcal{S}\right),\end{aligned}$$

where $H(X) := -\sum_{x\in\mathcal{X}} p(x)\log p(x)$ represents the entropy. The first term in the objective function promotes queries with a high uncertainty in the expected output, avoiding predictable answers which provide low information gain. The second term attempts to reduce uncertainty caused by intrinsic noise, such as when labeling neutral words like "table." In such cases, the uncertainty mostly stems from human labeling inconsistencies rather than a lack of exploration.

Empirically, we observe that actively selecting the question does not significantly impact performance. Therefore, to speed up computations, we select the question q_t uniformly at random. However, actively selecting the word set does significantly improve the performance. Given the combinatorially large number of possible sets $\binom{|\mathcal{X}|}{|\mathcal{S}|}$ over which to maximize, exhaustive optimization is computationally prohibitive. To mitigate the computational burden, we greedily aggregate a single word

$$\begin{aligned}\mathbf{s} = \operatorname*{argmax}_{\mathbf{s}\in\mathcal{X}} \; & H\left(\frac{1}{N}\sum_{n=1}^{N}\left[\mathbf{x}_t, y_t\middle|\widehat{\boldsymbol{\theta}}_n, q_t, \{\mathcal{S}, \mathbf{s}\}\right]\right) \\ & - \frac{1}{N}\sum_{n=1}^{N} H\left(\mathbf{x}_t, y_t\middle|\widehat{\boldsymbol{\theta}}_n, q_t, \{\mathcal{S}, \mathbf{s}\}\right) \quad (7)\end{aligned}$$

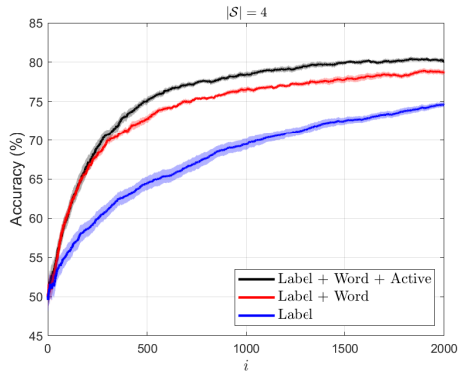
to the set until we reach size $|\mathcal{S}|$. Lines 5 through 10 in Algorithm 2 summarize the implementation of the active learning heuristic.

IV. EMPIRICAL RESULTS

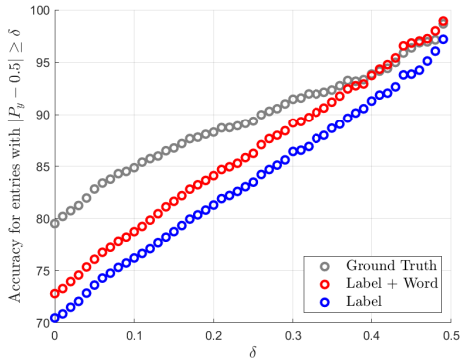
We empirically validate Algorithm 2. We use the list of most frequent words in the decade of the 2000s[1] [18]. For every word w, we simulate the implicit human score by sampling from $\mathcal{N}(\mu_w, \sigma_w^2)$, where μ_w and σ_w^2 are the mean and variance of the valence score as given by the dataset [18]. This simulation approach allows us to approximate a realistic distribution of human sentiment scores for each word, capturing the subjectivity between persons in valence assessments.

Given a word w with embedding $\mathbf{x}_w$ and a classifier estimator $\widetilde{\boldsymbol{\theta}}$, we claim the prediction $\mathrm{sign}(\widetilde{\boldsymbol{\theta}}^T\mathbf{x}_w)$ is accurate when it matches its label $\mathrm{sign}(\mu_w)$. To measure the estimator accuracy, we consider the words with $|P[Y_w = 1] - 0.5| = |\int_0^\infty f_{\mathcal{N}}(\mu_w, \sigma_w^2) - 0.5| \leq 0.1$, where $f_{\mathcal{N}}$ represents the probability density function of a normal distribution. This range is selected to avoid words that have a completely neutral score, like "branch" or "mouth." We focus on words with stronger sentiment scores, for which the prediction accuracy of our algorithm can be most meaningfully assessed.

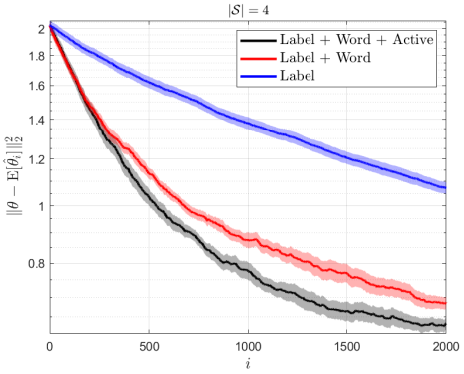
[1] https://nlp.stanford.edu/projects/socialsent/



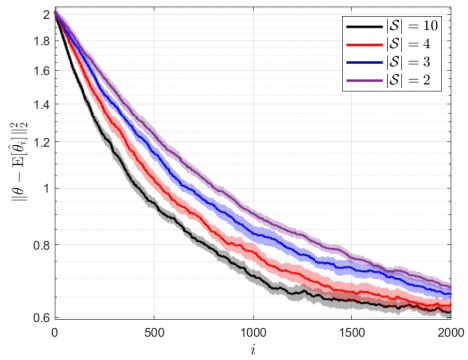
a) Accuracy vs. iteration for $|\mathcal{S}| = 4$



b) Accuracy vs. neutrality gap for $|\mathcal{S}| = 4$



c) Mean-Square Error (MSE) vs. iteration for $|\mathcal{S}| = 4$



d) $|\mathcal{S}|$ Comparison

Fig. 3: Performance of Algorithm 2 with human data. All configurations are run with 10 different random initializations. The solid lines represent the mean of among those experiments, while the shaded areas represent the standard error. Adding word selection to the queries together with actively selecting the word set reduces the number of iterations needed to achieve a good performance. The larger the word set, the greater the performance improvement is.

Figure 3a shows how the classification accuracy evolves with the number of queries. While the estimator accuracy increases by only asking the human for labels, incorporating word selection increases the estimator accuracy much faster. For example, to achieve an accuracy of 75% we would need more than 2000 iterations when relying solely on labels, while we only need around 700 iterations with q_{pos} and q_{neg}. This represents a notable reduction of 65% in the number of iterations needed, which we argue compensates for the additional effort required in word selection. We further improve the performance when we select the words in $\mathcal{S}$ actively; to achieve the same 75% accuracy, only 500 queries are necessary. This highlights the effectiveness of the method in reducing sample complexity.

We compare the estimated classifier after 2000 iterations of Algorithm 2 to the MMSE classifier, the ground truth classifier. As Figure 3b shows, the accuracy of the estimated classifier and the ground truth is similar when evaluating words at the polar ends of the valence spectrum, with high δ. The accuracy gap increases as words become more neutral. These words, which also present a challenge for human annotators, are typically considered of lesser priority in the realm of sentiment classification tasks due to their ambiguous nature [30]. Our approach reduces the accuracy gap to the ground truth classifier across the whole valence spectrum, compared to the traditional approach of only querying for the label.

In addition to the accuracy insights, Figure 3c depicts the evolution of the distance of the estimator to the ground truth as more queries are collected. Consistent with the previously discussed trends, we observe an advantage when incorporating word selection and active learning strategies. Algorithm 2 facilitates a faster reduction in MSE, evidencing not just an improvement in superficial accuracy metrics but also a genuine enhancement in the estimator alignment with the ground truth.

Figures 3a, 3b and 3c analyze the performance of Algorithm 2 for $|\mathcal{S}| = 4$. In Figure 3d we compare the effect of the size of the word set. The larger the word set the human chooses from, the more information the word selection has, thus the faster the classifier is learnt.

Collectively, these results confirm the efficiency of Algorithm 2, showcasing its applicability to the valence classification task. The integration of word selection alongside traditional label requests allows for more nuanced and informative responses, enhancing the classifier's learning process. We demonstrate a significant reduction in the number of iterations needed to achieve high classification accuracy.

ACKNOWLEDGEMENTS

We sincerely thank the reviewers for their valuable feedback. This work was supported by the Rafael del Pino Foundation.

REFERENCES

[1] S. Casper, X. Davies, C. Shi, T. K. Gilbert, J. Scheurer, J. Rando, R. Freedman, T. Korbak, D. Lindner, P. Freire, T. Wang, S. Marks, C.-R. Segerie, M. Carroll, A. Peng, P. Christoffersen, M. Damani, S. Slocum, U. Anwar, A. Siththaranjan, M. Nadeau, E. J. Michaud, J. Pfau, D. Krasheninnikov, X. Chen, L. Langosco, P. Hase, E. Bıyık, A. Dragan, D. Krueger, D. Sadigh, and D. Hadfield-Menell, "Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback," no. 3, 7 2023.

[2] Y. Liu, "Active learning with support vector machine applied to gene expression data for cancer classification," *Journal of Chemical Information and Computer Sciences*, vol. 44, no. 6, pp. 1936–1941, 2004.

[3] C. J. Cai, E. Reif, N. Hegde, J. Hipp, B. Kim, D. Smilkov, M. Wattenberg, F. Viegas, G. S. Corrado, M. C. Stumpe, and M. Terry, "Human-centered tools for coping with imperfect algorithms during medical decision-making," in *Proc. of Conference on Human Factors in Computing Systems*. ACM, 2019, p. 14.

[4] P. F. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," in *Advances in Neural Information Processing Systems*, vol. 2017-Decem, 2017, pp. 4300–4308.

[5] J. Fu, A. Korattikara, S. Levine, and S. Guadarrama, "From language to goals: Inverse reinforcement learning for vision-based instruction following," *International Conference on Learning Representations*, pp. 1–14, 2019.

[6] J. Zhu and E. Hovy, "Active Learning for Word Sense Disambiguation with Methods for Addressing the Class Imbalance Problem," in *Proc. of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, J. Eisner, Ed. Prague, Czech Republic: Association for Computational Linguistics, 2007, pp. 783–790.

[7] T. Nasukawa and J. Yi, "Sentiment analysis: Capturing favorability using natural language processing," *Proc. of International Conference on Knowledge Capture*, pp. 70–77, 2003.

[8] M. R. R. Rana, A. Nawaz, and J. Iqbal, "A survey on sentiment classification algorithms, challenges and applications," *Acta Universitatis Sapientiae, Informatica*, vol. 10, no. 1, pp. 58–72, 8 2018.

[9] B. Settles, "Active Learning Literature Survey," University of Wisconsin, Madison, Tech. Rep., 2009.

[10] P. Donmez and J. G. Carbonell, "Proactive learning: Cost-sensitive active learning with multiple imperfect oracles," *Proc. of International Conference on Information and Knowledge Management*, pp. 629–638, 2008.

[11] G. H. Canal, A. K. Massimino, M. A. Davenport, and C. J. Rozell, "Active embedding search via noisy paired comparisons," *Proc. of International Conference on Machine Learning*, vol. 2019-June, pp. 1493–1512, 2019.

[12] S. Li, S. Ju, G. Zhou, and X. Li, "Active learning for imbalanced sentiment classification," in *Proc. of Conference on Empirical Methods in Natural Language*, no. July, 2012, pp. 139–148.

[13] K. Lakshmi Devi, , P. Subathra, , and P. N. Kumar, "Performance Evaluation of Sentiment Classification Using Query Strategies in a Pool Based Active Learning Scenario," in *Advances in Intelligent Systems and Computing*, M. Senthilkumar, V. Ramasamy, S. Sheen, V. A. Bonato, and L. Batten, Eds. Singapore: Springer, 2015, pp. 65–75.

[14] B. Settles, "Active Learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 6, no. 1, pp. 1–114, 6 2012.

[15] S. Shekhar, M. Ghavamzadeh, and T. Javidi, "Active Learning for Classification With Abstention," *Journal on Selected Areas in Information Theory*, vol. 2, no. 2, pp. 705–719, 6 2021.

[16] Z. Ashktorab, M. Desmond, J. Andres, M. Muller, N. N. Joshi, M. Brachman, A. Sharma, K. Brimijoin, Q. Pan, C. T. Wolf, E. Duesterwald, C. Dugan, W. Geyer, and D. Reimer, "AI-Assisted Human Labeling: Batching for Efficiency without Overreliance," *Proc. of Human-Computer Interaction*, vol. 5, no. CSCW1, pp. 1–27, 2021.

[17] S. M. Mohammad, "Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words," in *Proc. of Annual Meeting of the Association for Computational Linguistics*, I. Gurevych and Y. Miyao, Eds., vol. 1. Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 174–184.

[18] W. L. Hamilton, K. Clark, J. Leskovec, and D. Jurafsky, "Inducing domain-specific sentiment lexicons from unlabeled corpora," in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 595–605.

[19] R. A. Bradley and M. E. Terry, "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons," *Biometrika*, vol. 39, no. 3/4, p. 324, 12 1952.

[20] Q. Liu, H. Huang, Y. Gao, X. Wei, Y. Tian, and L. Liu, "Task-oriented word embedding for text classification," in *Proc. of International Conference on Computational Linguistics*, E. M. Bender, L. Derczynski, and P. Isabelle, Eds. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018, pp. 2023–2032.

[21] M. Artetxe, G. Labaka, and E. Agirre, "Learning principled bilingual mappings of word embeddings while preserving monolingual invariance," in *Proc. of Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2016, pp. 2289–2294.

[22] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of Research in Personality*, vol. 11, no. 3, pp. 273–294, 9 1977.

[23] K. E. Train, *Discrete Choice Methods with Simulation*. Cambridge: Cambridge University Press, 2003, vol. 9780521816.

[24] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley, 9 2005.

[25] R. Ranganath, S. Gerrish, and D. M. Blei, "Black box variational inference," in *Journal of Machine Learning Research*, vol. 33, 2014, pp. 814–822.

[26] T. S. Jaakkola and M. I. Jordan, "Bayesian parameter estimation via variational methods," *Statistics and Computing*, vol. 10, pp. 25–37, 2000.

[27] Kevin P. Murphy, *Probabilistic Machine Learning Advanced Topics*. MIT Press, 2022.

[28] M. Braun and J. McAuliffe, "Variational inference for large-scale models of discrete choice," *Journal of the American Statistical Association*, vol. 105, no. 489, pp. 324–335, 12 2007.

[29] A. Kachites McCallum and K. Nigam, "Employing EM and Pool-Based Active Learning for Text Classification," in *Proc. of International Conference on Machine Learning*, J. Shavlik W., Ed. Morgan Kaufmann, 1998, pp. 350–358.

[30] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proc. of Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 2005, pp. 347–354.