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

Belen Martin-Urcelay\textsuperscript{1}, Christopher J. Rozell\textsuperscript{1}, Matthieu R. Bloch\textsuperscript{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 with theoretical analysis, providing sample complexity bounds. We also perform experiments with human data, demonstrating the effectiveness of our method in improving the accuracy of binary sentiment word classification.

I. INTRODUCTION

While humans are adept at classification, articulating the underlying rules that lead to these classifications remains elusive \cite{1}, leading to a gap between implicit human knowledge and explicit classification rules of learning algorithms. This gap makes it challenging to gather the knowledge of humans in a learning algorithm, and to teach machines how to make decisions informed by human experts, which is useful in a wide range of applications, including medical diagnosis \cite{2}, \cite{3}, robotics \cite{4}, \cite{5}, and word sense disambiguation \cite{6}.

We focus here on the binary sentiment word classification task \cite{7}. While humans can instinctively label words according to their connotation as positive (e.g., healthy) or negative (e.g., scary) \cite{8}, articulating the defining attributes of each category proves challenging for most. The transfer of implicit knowledge from humans to learning algorithms is further complicated by the fact that the information transferred is only useful if it is both comprehensible for humans and algorithms. As a result, the role of humans is often limited to serving as oracles that label data points \cite{9}, \cite{10}, \cite{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 \cite{12}, \cite{13} incorporate active learning \cite{14}, \cite{15} to select the most informative queries for human annotation. However, because of the limited role of humans only providing labels, the amount of information obtained from each query is low \cite{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 is required.

\textsuperscript{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

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 a word’s perceived sentiment and the word’s distance to the classifier. This model enhances the understanding and prediction of human word selection behaviours.
- 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 validated our approach theoretically and empirically. We provide bounds for the expected stopping time, and we evaluate the performance of the algorithm through experiments with human data.

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 $\theta \in \mathbb{R}^d$ that correctly labels words $x \in \mathcal{X} \subseteq \mathbb{R}^d$ as positive or negative according to $y = \operatorname{sign}(\theta^T x)$. An established method for learning this implicit knowledge, is to use humans as oracles within the framework of the Bradley-Torrey model \cite{19}. This model encapsulates the probability of a word being labeled as positive ($y = 1$) in the form of a logistic function

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

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 \cite{20} tailor embedding spaces to specific 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 as we move farther from the hyperplane in opposite directions to find words more positive or negative respectively. Heuristically, the intuition holds true for several lexicons in a preexisting word embedding [21]. Previous influential 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 their embeddings’ distance to the classifier. To formalize this observation, we introduce Assumption II.1.

Assumption II.1. The implicit sentiment score associated with a word \( x \) by a human is given by

\[
\text{score}(x) = ax^T \theta + b + \epsilon,
\]

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

Consider the question \( q_{\text{pos}} = \text{“Select and label the word you find most positive”} \); in that case, we may model the probability that the word \( x_i \) is selected from a word set \( S = \{x_t\}_{t=0} \) as

\[
\mathbb{P}[x_i | S, \theta, q_{\text{pos}}] = \mathbb{P}[\text{score}(x_i) > \text{score}(x_j), \forall j \neq i]
= \mathbb{P}[\epsilon_i - \epsilon_j < a(x_i - x_j)^T \theta, \forall j \neq i]
= \exp \left( \frac{a}{\sigma} x_i^T \theta \right) / \sum_{x \in S} \exp \left( \frac{a}{\sigma} x^T \theta \right),
\]

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

\[
\mathbb{P}[x_i | S, \theta, q_{\text{neg}}] = \exp \left( -\frac{a}{\sigma} x_i^T \theta \right) / \sum_{x \in S} \exp \left( -\frac{a}{\sigma} x^T \theta \right).
\]

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

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 \( q \in Q = \{q_{\text{pos}}, q_{\text{neg}}\} \) and word set \( S \)) such that the expected human’s response is as informative as possible, i.e., the \( q, S \) maximize the mutual information between the underlying classifier and the human response. This task is managed by the query
Algorithm 1. Ideal Human-in-the-Loop Learning

1: Input: $\mathcal{X}$, $Q$, $|S|$, $\mu_0$, $\Sigma_0$, $\epsilon^d$, $N$
2: $t = 0$
3: while $|\Sigma_{t}| < \epsilon^d$ do
4: $q_t, S_t \leftarrow \text{argmax } \mathbb{E}_{x, y, q, S} \left[ I(\theta; x, y, q, S, \mathcal{F}_{t-1}) \right]$
5: $x_t, y_t \leftarrow \text{human’s response to the query}$
6: $\mathbb{P}[\theta|\mathcal{F}_t] \propto \mathbb{P} [x_t|\theta, q_t, S_t] \mathbb{P} [y_t|x_t, \theta] \mathbb{P}[\theta|\mathcal{F}_{t-1}^{-1}]$
7: $t = t + 1$
8: end while

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

Algorithm 2. Approximate Human-in-the-Loop Learning

1: Input: $\mathcal{X}$, $Q$, $|S|$, $\mu_0$, $\Sigma_0$, $\epsilon^d$, $N$
2: $t = 1$
3: while $|\Sigma_{t}| < \epsilon^d$ do
4: $q_t, S_t \leftarrow \text{sample uniformly from } \{q_{pos}, q_{neg}\}$
5: $S_t \leftarrow \{\}$
6: for $i = 1, 2, \ldots, |S|$ do
7: $\{\theta_n\}_{n=1}^{N} \leftarrow \text{sample i.i.d. from } \mathcal{N}(\mu_{t-1}, \Sigma_{t-1})$
8: $s \leftarrow \text{select word from dictionary with Eq. (7)}$
9: $S_t \leftarrow \{S_t, s\}$
10: end for
11: $x_t, y_t \leftarrow \text{human’s response to the query}$
12: $\mu_t, \Sigma_t \leftarrow \mu_{t-1}, \Sigma_{t-1}$
13: while $\mu_t, \Sigma_t$ not converged do
14: while $\mu_t, \Sigma_t$ not converged do
15: $\xi = w^2 x^T \Sigma t x + w^2 (x^T \mu_t)^2$
16: $\Sigma^{-1} = \Sigma_{t-1}^{-1} + 2 \tanh(\xi/2) w^2 x x^T$
17: $\mu_t = \Sigma_t^{-1} [\Sigma_{t-1}^{-1} \mu_{t-1} + (y - \frac{1}{2}) w x]$
18: end while
19: $\mu_t, \Sigma_t \leftarrow \text{update according to the selected word with Equation (6)}$
20: end while
21: $t = t + 1$
22: end while

Approximation of Belief Update

Line 6 of Algorithm 1 requires us to find the posterior

$$
\mathbb{P}[\theta|\mathcal{F}_t] = \frac{\mathbb{P}[x_t|\theta, S_t, q_t] \mathbb{P}[y_t|x_t, \theta]}{\mathbb{P}[x_t|y_t, S_t, q_t, \mathcal{F}_{t-1}^{-1}]} \mathbb{P}[\theta|\mathcal{F}_{t-1}^{-1}].
$$

Unfortunately, the likelihood functions are not conjugates of the prior, so no analytical closed form expression exists to compute the posterior. Although, Black Box Variational Inference (BBVI) [24] is commonly used to approximate the posterior, our closed form derivation of the variational updates for this setting, which we describe next, provides a lower variance approximation.

We approximate the classifier’s density function as a Gaussian distribution $\theta \sim \mathcal{N}(\mu, \Sigma)$. We may then compute the posterior mean and variance given a word label by a variational approach.

We look for the variational distribution $q(\theta) \sim \mathcal{N}(\mu_q, \Sigma_q)$ 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) [26] such that

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

where $x_s$ is the word selected by the human, and the prior is $p(\theta) \sim \mathcal{N}(\mu_p, \Sigma_p)$. The first term in Equation (5) is the KL divergence between two Gaussian distributions

$$
\text{KL}(q||p) = \frac{1}{2} \left[ \log \frac{\sqrt{\Sigma_p}}{\sqrt{\Sigma_q}} - d + \mu_q^T \Sigma_p^{-1} \mu_q + \mu_p^T \Sigma_p^{-1} \mu_p \right]
$$

$$
- \mu_q^T \Sigma_p^{-1} \mu_p + \frac{1}{2} \log \left( \Sigma_p^{-1} \Sigma_q \right).
$$

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

$$
\mathbb{E}_{\theta \sim q} \left[ K x^T \theta \right] = K x^T \mu_q.
$$
The third term in Equation (5) has no closed-form solution, but following [27], we apply Jensen’s inequality to obtain a lower bound

\[
\mathbb{E}_{\theta \sim q}\left[ \log \prod_{j=1}^{|S|} \exp \left( K x_j^T \theta \right) \right] \\
\geq \log \prod_{j=1}^{|S|} \exp \left( K x_j^T \mu_q + 0.5x_j^T \Sigma_q x_j \right).
\]

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

\[
\{\mu, \Sigma\} = \arg\min_{\mu_q, \Sigma_q} KL(q||p) - K x_x^T \mu_q \\
+ \log \prod_{j=1}^{|S|} \exp \left( K x_j^T \mu_q + 0.5x_j^T \Sigma_q x_j \right). \tag{6}
\]

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 Equation (6). We repeat both updates until convergence.

**Active Learning Heuristic for Query Selection**

As line 4 of Algorithm 1 indicates, we want to select the query that provides the most information on expectation. This requires us to compute the posterior over every possible word set, query, label and word selection. Even using the belief approximation, there are combinatorially many options to compare, which makes this approach computationally intractable.

We propose an approximation based on query by committee [28]. At each iteration \(t\), we sample \(N\) particles \(\tilde{\theta}_n \sim \mathcal{P} \{\theta | F_t\}\). We maximize the disagreement between the prediction of each particle and the mean prediction among all particles as

\[
\{q, S_t\} = \arg\max_{q \in \mathcal{Q}, S \in \{X\}^d} H \left( \frac{1}{N} \sum_{n=1}^{N} \left[ x_t, y_t | \tilde{\theta}_n, q, S \right] \right) - \frac{1}{N} \sum_{n=1}^{N} H \left( x_t, y_t | \tilde{\theta}_n, q, S \right).
\]

The first term in the objective function promotes queries with a high uncertainty of the expected output, which avoids queries for which the answer is predictable and thus not very informative. The second term attempts to minimize uncertainty due to intrinsic noise, for example, discouraging asking labels of neutral words such as “table” for which the uncertainty mostly comes from the labelling noise from humans, and not from a lack of exploration.

| Assumption IV.2 | Word and label selected are independent of the history given the classifier, i.e., \(p(x_t, y_t | \theta, q_t, S_t, F_{t-1}) = p(x_t, y_t | \theta, q_t, S_t)\). |
|-----------------|------------------------------------------------------------------|
| Assumption IV.3 | The information gain of the classifier given a human answer is lower bounded by a constant \(L \in \mathbb{R}\), i.e., \(I(\theta; X_t, Y_t | F_{t-1}) \geq L > 0\). |

**Assumption IV.4** The prior distribution of the classifier \(p_0\) is uniform over a hypercube \([-M, M]^d\).

Our main result is the following:

**Theorem IV.5.** Given the stopping time \(T \leq \min\{t : |\Sigma_{\theta|F_t}|^{1/d} < \epsilon\}\), under Assumptions II.1 through IV.4, the expected stopping time for Algorithm 1 is bounded as

\[
\frac{d}{2} \log_2 \frac{4M^2}{2\pi \epsilon^2} \leq \mathbb{E}[T] \leq \frac{d}{2L} \log_2 \frac{e^2 L^2 M^2}{2\sqrt{2(d+2)} \epsilon} - 1.
\]
Proof: For the lower bound, we note that Algorithm 1 selects the query deterministically as a function of its latest classifier estimator, i.e., \( I(\theta; q_t, S_t|F_i) = 0 \), so
\[
\begin{align*}
  h(\theta|F_i) &= h(\theta|F_{i-1}) - I(\theta; x_t, y_t, q_t, S_t|F_i) \\
  &= h(\theta|F_{i-1}) - I(\theta; x_t, y_t|q_t, S_t, F_i) \\
  &\geq h(\theta|F_{i-1}) - \max I(\theta; x_t, y_t|q_t, S_t, F_i) \\
  &= h(\theta|F_{i-1}) - \log 2|S| \\
  &\geq h(\theta|F_0) - t \log 2|S| \\
  &= d \log 2 M - t \log 2|S|. 
\end{align*}
\]

because the query that is most informative on expectation occurs when every word-label combination has the same probability of \( \frac{1}{2^{|S|}} \). The last inequality comes from Assumption IV.4. This implies
\[
E[T_i] \geq \frac{d \log 2 M - h(\theta|F_{T_i})}{\log 2|S|}. \tag{9}
\]

As shown in Lemma VI.1, \( h(\theta|F_i) \) is log-concave (LCC) so we invoke [11, Lemma 3.1.]
\[
h(\theta|F_{T_i}) \leq \frac{d}{2} \log 2 \pi e |\Sigma_{\theta} x_{F_{T_i}}|^{1/d} \leq \frac{d}{2} \log 2 \pi e. \tag{10}
\]

Combining Equation (9) and (10)
\[
E[T_i] \geq \frac{d \log 2 M^2}{\log 2|S|}. \tag{11}
\]

To obtain the upperbound, we define the random variable \( U_i := Z_i - t \), where \( Z_i = -h(\theta|F_i) \). From Lemma VI.4 we know \( U_i \) is a submartingale and it fulfills the conditions of the optional stopping theorem [11], so
\[
\frac{E[Z_{T_i}]}{L} - E[T_i] \geq \frac{E[Z_0]}{L} - E[0] \\
\iff \frac{E[Z_{T_i}] - E[Z_0]}{L} \geq E[T_i]. \tag{11}
\]

By Assumption IV.4 the initial entropy is
\[
Z_0 = -h(\theta|F_0) = -d \log 2 M. \tag{12}
\]

From Assumption IV.3
\[
E[Z_{T_i}] \leq E[Z_{T_{i-1}}] - L. \tag{13}
\]

We invoke again [11, Lemma 3.1.] to obtain
\[
Z_{T_{i-1}} = -h(\theta|F_{T_{i-1}}) \leq -\frac{d}{2} \log 2 |\Sigma_{\theta|F_{T_{i-1}}}|^{1/d} \cdot e^{4d^2/(4\sqrt{2}(d+2))} \\
\leq -\frac{d}{2} \log 2 \sqrt{8d(d+2)e} \cdot e^{4d^2/(4\sqrt{2}(d+2))} \\
= \frac{d}{2} \log 2 \frac{e^{4d^2}}{8\sqrt{2}(d+2)e}. \tag{14}
\]

Substituting Equations (12), (13) and (14) into Equation (11), we obtain the upper bound of the expectation.
\[
\begin{align*}
E[T_i] &\leq \frac{1}{L} (E[Z_{T_{i-1}}] - L + d \log 2 M) \\
&\leq \frac{d}{2L} \left( \log 2 \frac{e^{4d^2}}{8\sqrt{2}(d+2)e} + \log 2 4M^2 \right) - 1 \\
&\leq \frac{d}{2L} \log 2 \frac{e^{4d^2}}{2\sqrt{2}(d+2)e} - 1. 
\end{align*}
\]

Theorem IV.5 extends the upper bound in [11] to non-equiprobable queries. The parameter \( M \) appears because we allow for more freedom in the prior. The more constrained the prior is, lower \( M \), the less interactions we need. Because of the more informative queries, the denominator \( \log 2|S| \) appears in the lower bound. This generalizes the lower bound of [11] beyond binary queries. A lower stopping time may be possible as the word set size increases. Although the upper bound does not yet capture this behavior, we empirically corroborate the dependence of the stopping time on the word set size, as suggested by the lower bound, in the next section.

V. 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 \( N(\mu_w, \sigma_w^2) \), where \( \mu_w \) and \( \sigma_w^2 \) 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 \( x_w \), we claim the prediction \( \text{sign}(\theta^T x_w) \) is accurate when it matches its label \( \text{sign}(\mu_w) \). To measure the predictor’s \( \theta \) accuracy, we consider the words with \( |P[Y_w = 1] - 0.5| = \int_0^{\infty} f_N(\mu_w, \sigma_w^2) - 0.5 \leq 0.1 \), where \( f_N \) 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.

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 \( q_{\text{pos}} \) and \( q_{\text{neg}} \). 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 \( S \) actively; to achieve the same 75% accuracy, only 500

\(^1\)https://nlp.stanford.edu/projects/socialsent/
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.

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 $\delta$. 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 [29]. 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 $|S| = 4$. In Figure 3d we compare the effect of the size of the word set. As the lower bound in Theorem IV.5 suggests, the larger the word set the faster the classifier is learnt.

Our results confirm the efficiency of Algorithm 2, showcasing its applicability to the valence classification task. To facilitate further exploration and validation by the research community, we provide the code for replicating our experiments².

REFERENCES


²https://github.com/BelenMU/HiTL-SentimentClassify/
Lemma VI.1. The posterior distribution of the classifier given the history, i.e., \( p(\theta | F_i) \) is LCC.

Proof:

\[
p_i(\theta) := p(\theta | F_i) = p(\theta|x_i, y_i, q_i, S_i, F_{i-1})
\]

\[
= \frac{p(x_i, y_i, q_i, S_i, \theta, F_{i-1}) p(\theta | q_i, S_i, F_{i-1})}{p(x_i, y_i, q_i, S_i, F_{i-1})}
\]

\[
= \frac{p(x_i, y_i, q_i, S_i, \theta)}{p(x_i, y_i | F_{i-1})} p(\theta | F_{i-1})
\]

\[
= \sum_{j=1}^{3} \frac{p(x_j, y_j | q_j, S_j, \theta) p_0(\theta)}{p(F_j)}
\]

where

1) follows from Bayes theorem,
2) follows from Assumption IV.1, and because given the past history the query and word set are selected deterministically by the algorithm,
3) analysing the recursion,
4) combining Bayes, the law of total probability and Assumption IV.2.

The likelihood of the word and label selected, \( p(x_j, q_j, S_j, \theta) \) and \( p(y_j | q_j, x_j, \theta) \), are given by the softmax function and the logistic function respectively, both of which are LCC. As for Assumption IV.4, the prior \( p_0(\theta) \) is sampled from a uniform distribution, which is also LCC. The product of LCC functions is also LCC, thus the posterior \( p_i(\theta) \) is LCC.

Lemma VI.2. The likelihood of any word and label pair \( x \in X \) and \( y \in \{-1,1\} \) is lower bounded as

\[
\min_{x, y, S, \theta, q} P[x, y | S, \theta, q] \geq 2^{-\gamma} > 0,
\]
with $\gamma_L := \log_2 \left( \frac{\exp\left(-\frac{|\gamma|}{2} \right)}{\exp\left(-\frac{|\gamma|}{2} M \sqrt{p \sigma^2} \right) + (|S| - 1) \exp\left(|\gamma| M \sqrt{p \sigma^2} \right)} \right)$
and $\gamma_y = \frac{1}{1 + \exp(w M \sqrt{p \sigma^2})}$.

Proof: As per construction, all word embeddings are bounded, i.e., $|x|^2 \leq P_x$. From Assumption IV.4, we know the norm of the classifier is bounded by $\|\theta\|^2 \leq dM^2$. We leverage these constraints to bound the likelihood.

Note that

$$\exp\left(-\frac{|\gamma|}{2} \right) \leq \exp(kx^T \theta) \leq \exp\left(\frac{a}{\sigma} M \sqrt{p \sigma^2} \right)$$

Therefore,

$$\min_{x, S, \theta, q} \mathbb{P}[x, S, \theta, q]$$

$$\geq \frac{\exp\left(-\frac{|\gamma|}{2} \right) M \sqrt{p \sigma^2}}{\exp\left(-\frac{|\gamma|}{2} M \sqrt{p \sigma^2} \right) + (|S| - 1) \exp\left(|\gamma| M \sqrt{p \sigma^2} \right)} \gamma_y$$

In a similar fashion, we use Assumption IV.2 to bound

$$\min_{x, y, \theta, q, S} \mathbb{P}[y|x, \theta, q, S]$$

$$= \min_{x, \theta} \min_{y \in (-1, 1)} \left( 1 + \exp\left( yw^T x \right) \right)$$

$$\geq \frac{1}{1 + \exp(w M \sqrt{p \sigma^2})} = \gamma_y$$

We conclude the proof by combining the bounds

$$\min_{x, S, y, \theta, q_i} \mathbb{P}[x, y|S, \theta, q_i]$$

$$\geq \exp\left(-\frac{|\gamma|}{2} M \sqrt{p \sigma^2} \right) \gamma_y$$

Lemma VI.3. The difference in entropy from one iteration to the next is bounded as

$$|h(\theta|F_i) - h(\theta|F_{i-1})| \leq \gamma < \infty,$$

with $\gamma = 8d + \frac{d}{2} \log_2 2\pi ed - \gamma_L$.

Proof: We extend [11, Lemma A.2.] to bound the entropy difference

$$-h(\theta|F_i) \leq \log_2 \mathbb{E}_{\theta|F_i} [p(\theta|F_i)]$$

$$\geq 2 \log_2 \mathbb{E}_{\theta|F_i} \left[ \frac{p(x_i, y_i|\theta, q_i, S_i, F_{i-1})}{p(x_i, y_i|q_i, S_i, F_{i-1})} p(\theta|F_{i-1}) \right]$$

$$\leq 3 \log_2 \mathbb{E}_{\theta|F_i} \left[ \frac{p(x_i, y_i|q_i, S_i, F_{i-1})}{p(\theta|F_{i-1})} \right]$$

$$\leq 4 \log_2 \left( \frac{1}{2} \right) + 8d + \frac{d}{2} \log_2 d$$

$$\leq 5 \log_2 \left( \frac{1}{2} \right) + \frac{1}{2} \log_2 (2\pi e)^d$$

$$\leq 6 \log_2 \left( \frac{1}{2} \right) + \frac{d}{2} \log_2 2\pi ed - \gamma_L$$

where the inequalities follow:

1) from Jensen’s inequality,
2) from Bayes Theorem and because the query is determined by the history,
3) because the likelihood is a valid probability distribution, thus it is at most 1, and the logarithm is monotonically increasing,
4) applying the bound in [30, Theorem 5.14] to the LCC isotropic density function $V = \Sigma_{\theta|F_i} \theta|F_i$,
5) adding and subtracting $\frac{1}{2} \log_2 (2\pi e)^d$.
6) from the maximum entropy distribution [31, Theorem 8.6.5.] and Lemma VI.2.

To bound the other direction, we recall the non-negativity of mutual information which implies $0 \leq h(\theta|F_{i-1}) - h(\theta|F_i)$. Combining the bounds

$$|h(\theta|F_{i-1}) - h(\theta|F_i)| \leq \gamma.$$

Lemma VI.4. The random variable $U_i := \frac{h(\theta|F_i)}{L} - i$ is a submartingale.

Proof: The expectation of $U_i$ given the previous values of the sequence is

$$\mathbb{E}[U_i|U^{i-1}] = \frac{\mathbb{E}[Z_i|Z^{i-1}] - i}{L} \geq \frac{\mathbb{E}[Z_{i-1}|Z^{i-1}] + L}{L} - i$$

$$= \frac{Z_{i-1} - (i - 1)}{L} = U_{i-1},$$

where $Z_i = -h(\theta|F_i)$. The first inequality follows from Assumption IV.3.

We bound the increment step per statement with Lemma VI.3,

$$|U_{i+1} - U_i| = \left| \frac{Z_{i+1}}{L} - i - 1 - \frac{Z_i}{L} + i \right|$$

$$\leq \frac{\gamma}{L} + 1.$$ (16)

We use Equation (16) to prove the absolute value is finite

$$|U_i| = \left| \sum_{j=1}^{i} U_j - U_{j-1} \right| \leq \sum_{j=1}^{i} |U_j - U_{j-1}|$$

$$\leq \frac{Z_0}{L} + i \gamma = \frac{\log_2 2|S|}{L} + i \gamma < \infty,$$ (17)

where the last inequality follows from Assumption IV.4.

Together Equations (15) and (17) prove $U_i$ is a submartingale.