Human-like Explanation for Text Classification With Limited Attention Supervision

Dongyu Zhang
Data Science Program
Worcester Polytechnic Institute
Worcester, USA
dzhang5@wpi.edu

Cansu Sen
CodaMetrix
Boston, USA
cansu@codametrix.com

Jidapa Thadajarassiri
Data Science Program
Worcester Polytechnic Institute
Worcester, USA
jthadajarassiri@wpi.edu

Thomas Hartvigsen
Data Science Program
Worcester Polytechnic Institute
Worcester, USA
twhartvigsen@wpi.edu

Xiangnan Kong
Computer Science Department
Worcester Polytechnic Institute
Worcester, USA
xkong@wpi.edu

Elke Rundensteiner
Computer Science Department
Worcester Polytechnic Institute
Worcester, USA
rundenst@wpi.edu

***Abstract*—Human-like explanation for text classification is essential for high-impact settings such as healthcare where human rationales are required to support specialists' decisions. Conventional approaches learn explanations using attention mechanisms to assign heavy weights to words that have a high impact on a model's prediction. However, such heavily-weighted words often do not reflect human intuition. To advance human rationale, recent studies propose to supervise attention mechanisms assuming access to a huge set of attention labels collected from humans, called *human attention maps* (HAMs). Unfortunately, acquiring such HAMs for a huge dataset is very tedious, error-prone, and expensive in practice. Thus, we propose the novel problem of text classification with *limited* human attention supervision. Specifically, we study the learning of *human-like attention weights* from a dataset in which all documents contain classification labels but only a few documents provide HAMs. To this end, we design a deep learning architecture, HELAS: Human-like Explanation with Limited Attention Supervision to adaptively learn attention weights that focus on words analogous to a human with very limited attention supervision. HELAS effectively unifies joint learning improving both tasks of text classification and human-like explanation even with only insufficient supervision labels for the latter task. Our experiments show that HELAS generates attention maps similar to real human annotations raising similarity scores up to 22% over state-of-the-art alternatives, even with as little as 2% of the documents having HAMs. It concurrently improves text classification by driving accuracy up to 19% over four state-of-the-art methods.**

***Index Terms*—Model Explainability, Text Classification, Joint Learning, Attention Mechanism**

I. Introduction

Motivation. Text classification is a crucial text mining task with broad applications including fake news detection [1], clinical diagnosis [2], and sentiment analysis [3]. With the availability of massive training corpora, several modern approaches [4]–[6] achieve impressive performance. Yet they are remain largely inapplicable in settings where explanations are required to support a decision. For example, a doctor must know on what information a diagnostic model relies before trusting its predictions. Attention-based models [7]–[9] can be used to acquire such explanations by learning to assign heavy weights to words that have a high impact on a model's prediction.

Recently, there is growing evidence that attention weights that look as if they were generated by humans lead to both better explanations and sometimes even improved classification [10], [11]. However, attention generated by conventional attention approaches are dissimilar to human rationales [11], [12]. Classic attention contradicts the ultimate goal of producing *explainable* models that allow human users to understand a model's rationale for a given prediction. Recent works [13]–[18] have begun to overcome this hurdle, enhancing *explanations* by encouraging them to be *human-like*, or resemble rationales provided by humans. This has been achieved by collecting additional attention labels and explicitly *supervising* the attention mechanism.

State-of-the-Art. Conventional approaches to supervised attention for text classification [10], [15], [16], [19] use hand-picked lists of *words-of-interest*, defined by a rule or by a domain expert, to serve as word-level attention labels. For training classifiers, attention weights that deviate from these lists are penalized. However, this fixed list of words-of-interest is used for all input text. Thus, this is not likely to naturally lead to human-like attention due to this rigidity. To overcome this, recent work [18] has turned to *human-gaze attention data* collected from large corpora as a static source to supervise attention methods. However, as this approach again is static using general knowledge as attention on unrelated sources of text; there is a disconnect between how human attention was initially recorded and the final classification task.

This work was done when Cansu Sen was at Worcester Polytechnic Institute. This work was supported by the NSF Division Information and Intelligent Systems (awards 1815866, 1852498, 1910880, and 1718310), the GAANN Fellowship in Computer and Information Sciences for AI from the U.S. Dept. of Education (P200A150306 and P200A180088), and the Royal Thai Government.

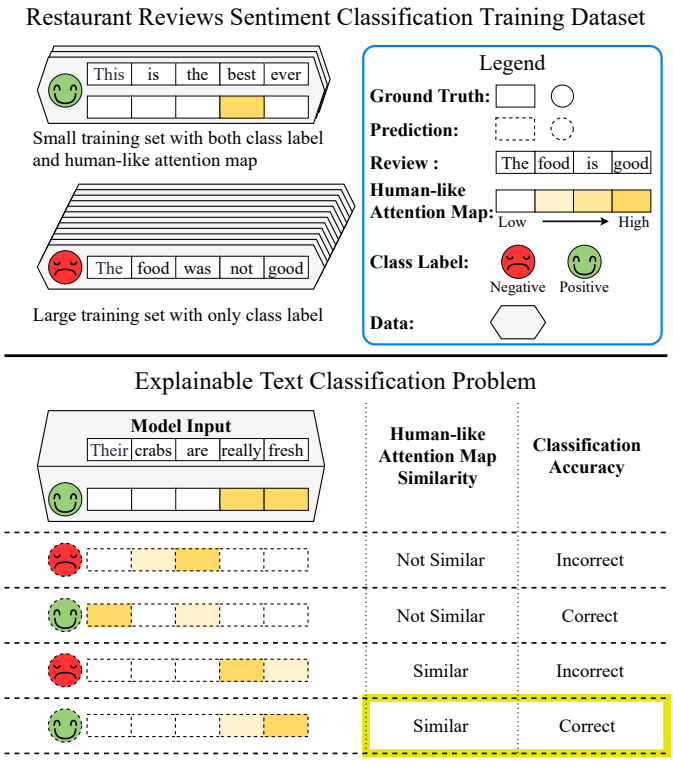


Fig. 1: Explainable text classification with limited human attention supervision. Given a corpus of documents, each with a document-level label for classification task while only a few with word-level labels (human attention maps) for supervising attention, the dual goal is to learn a model that classifies text documents accurately and generates human-like word attention maps.

Meanwhile, attention weights have also been learned directly from *human rationales* specific with each training input document [17], [20], indicating which set of words a human pays more or less attention to while they perform a classification task. Using such a set of fine-grained word-level attention labels collected for the task at hand, referred to as a *human attention map* (HAM), is an extremely promising approach for creating document-specific human-like attention mechanisms [11]. As proposed, these approaches require access to HAMs for every single training document along with each document's corresponding classification label. In practice, however, collecting HAMs is far more expensive than classification labels alone since it requires annotators to dedicate much more effort and time to consider every word in a huge dataset. This is even more impractical for highly-technical domains such as healthcare where expert annotators are rare and busy.

Problem Definition. In this work, we are the first to study the problem of *explainable text classification with limited human attention supervision*. This addresses the real-world case where access to HAMs is severely limited. As illustrated in Figure 1, assume we are given a set of training documents, each with one associated classification label. A very small proportion of these training documents also have fine-grained word-level labels (HAMs), indicating which words a human annotator found to be most relevant as they assigned the class label. Our goal is to train a model that simultaneously solves the text classification task accurately while predicting *human-like attention weights* that are similar to those that would be generated by a human for the given document.

Challenges. Text classification with limited human attention supervision is challenging for the following reasons.

• *Sensitivity of attention to changing contexts.* A word with high human attention in one document does not necessarily have high human attention in the other document. This implies that the attention weight for a word relies heavily on the context in which it appears. A successful attention method must effectively capture this reliance between context and human-like attention.

• *Conflict between human-like attention generation and text classification.* Our problem requires a model to assign specialized weights to individual words. However, every word contributes to the classification task. Therefore, unsupervised attention weights are often more distributed across a sentence than a HAM. A successful model must balance between the two contradictory objectives of human-likeness and classification accuracy.

• *Varying levels of supervision.* This problem has two tasks: classification and human-like attention generation. However, each of the two tasks has a different amount of labeled data — all data have classification labels, only some have human attention maps. A good solution must balance the feedback given from each task without overemphasizing the fully-labeled task.

Proposed Method. To handle these challenges, we propose the deep learning architecture, HELAS: Human-like Explanation with Limited Attention Supervision, which produces human-like attention values during text classification, even when very few human attention labels are available. HELAS processes input text in three phases : (1) HELAS encodes input text through a *text representation learner* into both dense vectors for each word and one vector for the whole document. This text representation learner is highly modular and can learn representations using many recent text models such as RNNs [21], [22] or BERT [5]. (2) The *human-like attention learner* in HELAS learns human-like attention weights for each word by both considering its individual impact on the classification task *and* by carefully incorporating its contextual information. This allows the learned attention mechanism to be adaptive to context, similar to a human annotator. (3) The *contextualized representation* collates the contextualized information learned according to the human-like attention learner with the overall text representation to consider *both* sources or information and perform the final classification. Thus, our approach succeeds to capture the unique contribution of each word in a given document and produce both human-like attention and accurate classifications.

HELAS is optimized using a joint loss function for the classification and human-like attention-learning tasks. We introduce a hyper-parameter into the loss function for striking

a balance between classification and attention supervision, resulting in one unified training objective. This newly defined loss handles the varying levels of supervision for both classification and attention supervision and thus allows HELAS to deliver accurate classifications and human-like attention weights simultaneously.

Contributions. Our contributions are as follows:

• We define the open problem of explainable text classification with limited human attention supervision, which is to develop a human-like explainable classifier when few HAMs are available.

• We propose the first solution to this problem, HELAS, which contains two key components: (1) a novel attention method, called human-like attention learner, that successfully learns human-like attention weights, adapting to different contexts, and (2) a custom contextualized representation that considers the impact of all words to make its final prediction.

• We propose a joint loss function for HELAS that balances the limited attention supervision and fully-supervised classification supervision, encouraging the model to generate more human-like attention values – even with very few HAMs.

• We demonstrate that even when HAMs are available for as little as 2% of the training data, HELAS still succeeds to generate human-like attention, achieving up to 22% increase in similarity compared to four state-of-the-art methods. HELAS also gets better performance on the classification task achieving significant (up to 19%) gains in accuracy.

II. Related Works

Supervised Attention Models. Attention supervision is used for NLP problems. In [23], [24], conventional alignment models are used to guide the attention module for language translation. [15] apply supervised attention method for event detection, namely, their model focuses on event information on both the word- and sentence-level. [16] introduce supervised attention for improving the accuracy of the semantic event recognition; namely, by deploying semantic word lists and dependency parsing trees [25] to guide the attention components. [18] propose a method to use estimated human attention derived from eye-tracking corpora to regularize attention functions for sequence classification tasks. While these works show supervised attention can improve accuracy, the forms of guidance adopted remain limited – none of the methods mentioned above get attention guidance via word-level human attention maps collected for the classification task.

[17] propose a model with an attention mechanism for text classification that jointly exploits document classification labels and sentence-level annotation labels. They assume that annotators explicitly mark sentences that support their overall document categorization for each document in the corpus. However, collecting fine-grained sentence-level or word-level annotation labels for all instances in a dataset can be costly and time-consuming. Moreover, in [17], training with each level of labels is split into two steps. It is time-consuming and sophisticated to train their model. Hence, it is worthy of exploring a method that can be trained efficiently with limited access to HAMs.

Model Explainability. Deep-learning models suffer from a lack of explainability, despite the need for explainable models in many domain settings. Thus, several studies in recent years attempt to make neural network models more explainable. Rationale-based methods are examples of this for NLP [26], [27]. In these works, the goal is to train a classification model and produce binary "rationales" to serve as human-like explanations of model predictions. However, while their direction is promising, their classification performance remains a drawback compared to recent attention-based approaches [11]. Also, these rationale-based architectures make classifications based on the selected "rationales", not the full text [26], [27]. So the information in these non-rationale words is missing during prediction.

Recent work in deep learning instead has begun to use attention mechanisms to attempt to bring interpretability to model predictions [7]–[9]. However, these works assess the produced attention maps solely qualitatively by visualizing a few hand-selected instances.

[11] approaches attention explainability from a human-centered perspective. They investigate the similarity between human attention and machine attention and interpret such similarity as a measurement of the model explainability. It indicates that it is intuitive to humans as it matches which words humans would rely on when making decisions. [11] makes a novel human attention map resource available to the community. Inspired by their approach, we now leverage human attention to explicitly train a model to concurrently produce the overall task prediction as well as the *human-like explanations* with the power of modern attention mechanisms.

III. METHODOLOGY

A. Problem Definition

In this paper, we study the problem of *explainable text classification with limited human attention supervision*. Given a set of N documents $\mathcal{I} = \{\mathcal{D}^1, \ldots, \mathcal{D}^N\}$, each document $\mathcal{D}^i$ consists of T words $\mathcal{D}^i = [w_1^i, \ldots, w_T^i]$, and a set of class labels $y^i = [y_1^i, \ldots, y_K^i]$, where K is the cardinality of y^i, $y_k^i \in \{0, 1\}$ and $\sum_{k=1}^{K} y_k^i = 1$. The *document classification* task is to parameterize a function $f_\theta(\cdot)$ that maps $\mathcal{D}^i \rightarrow y^i$, generalizing to unseen instances.

A *Human Attention Map* (HAM) is a vector of length T, $[\alpha_1, \ldots, \alpha_T]$, where each entry α_t indicates the degree of attention that a human pays to a corresponding word w_t in a document. HAM is a binary map collected from humans, *i.e.*, $\alpha_t = 1$ indicates that the corresponding word receives high attention while 0 shows low attention. A *Machine Attention Map* (MAM $= [\hat{\alpha}_1, \ldots, \hat{\alpha}_T]$) is a human-like attention map *predicted* by a neural network model, where $\hat{\alpha}_i \in [0, 1]$ indicates the probability of the corresponding word that would receive high attention from humans.

For each document $\mathcal{D}^i$, we are given a class label y^i. However, only a limited amount of documents have HAMs. One component of $f_\theta(\cdot)$ is an attention mechanism that aims

to output MAMs that are similar to HAMs. Our task is to jointly learn the function $f(\theta)$ while minimizing the difference between HAM^i and MAM^i for all documents $\mathcal{D}^i$, the latter task is named *human-like attention generation*. If conditioned perfectly, $f_\theta(\mathcal{D}^i) \rightarrow (\hat{y}^i, MAM^i)$ such that $\hat{y}^i = y^i$ and MAM^i = HAM^i.

For readability, we henceforth describe our method for a single document D^i, dropping i when it is unambiguous.

B. Proposed Method: HELAS

Our proposed deep learning architecture, HELAS: Human-like Explanation with Limited Attention Supervision, is depicted in Figure 2. HELAS consists of three major components: (1) The *text representation learner* encodes raw text to their numerical representations. This component can be any sequential deep learning architecture, such as RNNs [21], [22] or BERT [5]. The purpose of this layer is to encode the input document into a document representation and a sequence of word representations. (2) The *human-like attention learner* generates a MAM aimed to be similar to the given HAM. The attention mechanism determines the human-like attention weight for each word by the interrelation between word and sentence representations. (3) The *contextualized representation* utilizes the MAM from the human-like attention learner to enhance the context vector to estimate the class label, y, of a document.

1) Text Representations Learning: We focus our study on the two popular sequence modeling including RNNs [21], [22] and BERT [5] while HELAS can be, in practice, paired with any sequence-representation learning architectures.

HELAS-RNN. One common and powerful architecture for document classification is an RNN combined with an attention mechanism [7], [28]. Following this architecture, the HELAS-RNN model first utilizes an encoding layer to map words into real-valued vector representations where semantically-similar words are mapped close to one another. We use a pre-trained word embedding set ϕ for this mapping: $x_t = \phi w_t$. HELAS-RNN then employs a recurrent layer to embed vector representations of words into hidden states, processing words once at a time. In our experiments, we use both LSTM [21] and GRU [22] memory cells.

Assuming that Γ is the recurrence function (*e.g.*, LSTM or GRU) and x_t is the embedded t-th word from the document $\mathcal{D}$, HELAS-RNN is modeled as:

$$e_t = \Gamma(x_t, e_{t-1}) \tag{1}$$

where e_t is the hidden state. The final hidden state e_T is used as the document representation, defined as $r = e_T$.

HELAS-BERT. HELAS-BERT first employs a transformer architecture [4] to encode words, initialized with a pre-trained BERT model [5]. Following the standard practice in BERT-based architectures, the first word of the input is the special token '[CLS]'. '[SEP]' token is added to the end of the input sequence to denote the end. '[PAD]' token is used to pad the sequence in case the input sequence is shorter than the maximum input length supported by the BERT model. HELAS-

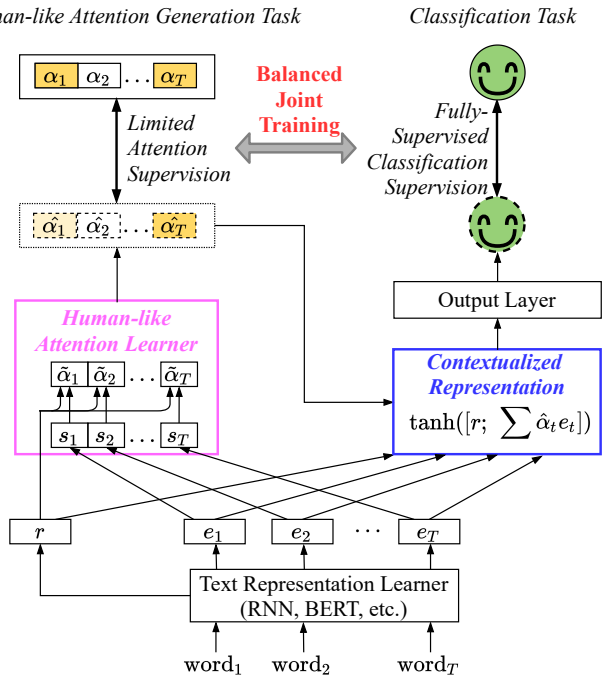


Fig. 2: Overall architecture of HELAS.

BERT generates two outputs. First is a sequence of learned word representations $[e_1, ...e_T]$ for each input word. Second is a vector representation r for the whole input document. This vector r corresponds to the output of the '[CLS]' token further processed by a linear layer and a $\tanh$ activation function.

$$[e_1, ...e_T], r = \text{BERT}([w_1, ...w_T]) \tag{2}$$

2) Human-Like Attention Generation: The goal is to generate attention scores to be as close as possible to human attention. This way, attention scores can be interpreted as human-like explanations for the final classification decision.

We hypothesize that the importance of each word relies heavily on its belonging document. Therefore, HELAS is designed to learn specialized attention function that is adaptable for different training corpus. The human-like attention learner learns MAMs as follows:

$$s_t = \tanh(W_e e_t + b_e) \tag{3}$$
$$\tilde{\alpha} = s_t^\top r \tag{4}$$
$$\hat{\alpha} = \text{sigmoid}(\tilde{\alpha}) \tag{5}$$

where W_e and b_e are weight matrix and bias term in the linear layer, which can be optimized during the training time. Here, we first encode new word representations s_t from e_t. s_t serves as a specialized representation of the importance of w_t, while e_t is still a general representation of the information contained in w_t. Then the raw attention score $\tilde{\alpha}$ is determined by such s_t and the document representation r in order to induce MAMs to capture more flexible relation between e_t and r. MAM = $[\hat{\alpha}_1, \ldots, \hat{\alpha}_T]$ is then utilized by the subsequent layers of the

HELAS. To this end, we take the binary cross entropy as the general loss of the attention at the word level.

$$J_a(\text{HAM}, \text{MAM}) = -(\sum_{t=1}^{T}(\alpha \log \hat{\alpha}_t + (1-\alpha_t)\log(1-\hat{\alpha}_t))). \tag{6}$$

This objective optimizes the model to assign human-like attention scores to every word. By providing word-level supervision to the document classification model, we are able to teach it to focus on the most relevant areas selected by humans and thereby improve the quality of document representations along with the overall performance.

It is worth noting that special tokens, such as '[CLS]', '[SEP]' and '[PAD]', are invisible to human annotators (if the text representation learner is BERT). Thus, their corresponding human attention weights are always set to zero. Also, the tokenizer used by the BERT model is WordPiece [29], which sometimes splits a word into several words. These generated words are then assigned with the same human attention score as the original word.

3) Document Classification: Using the MAMs, the learned word representations e_t and the document representation r generated by the text representation learner, the contextualized representation c is computed as follows:

$$c = \tanh([r; \sum_t \hat{\alpha}_t e_t]). \tag{7}$$

Unlike the previous works [17], [18] which use $\sum_t \hat{\alpha}_t e_t$ as the final text representation for classification, we concatenate document representation r and weighed sum of word representations to model a dense embedding for the document. For a given HAM, when $\alpha = 0$, it does not indicate that the corresponding word was completely ignored by humans. During training, the values of some $\hat{\alpha}$s could be very close to 0. Since r contains basic information of the whole document, the contextualized representation will consider every word when performing classification, even if some words' associated $\hat{\alpha} \approx 0$. Those words with higher $\hat{\alpha}$ values remain a higher impact on final prediction results.

The output layer uses c as follows:

$$d = \text{Dropout}(W_c c + b_c) \tag{8}$$

$$\hat{p}(y_k = 1|D) = \frac{\exp(W_d^{(k)} d + b_d)}{\sum_{k=1}^{K} \exp(W_d^{(k)} d + b_d)} \tag{9}$$

To further fuse r and $\sum_t \hat{\alpha}_t e_t$ together and reduce the risk of overfitting, we apply a linear transformation followed by a dropout layer in Equation 8. Here, W_c, b_c are weight matrix and bias term in the linear layer. After dropout, Equation 9 assigns a probability to each possible class, where W_d, b_d are weight matrix and bias term in the softmax function. We use the cross-entropy loss as the document classification objective function where $\hat{p}(y_k = 1|D)$ is the prediction and y the ground truth label.

$$J_c(y, \hat{y}) = -\sum_{k=1}^{K} y_k \log(\hat{p}(y_k = 1|D)). \tag{10}$$

4) Joint Training of HELAS: In HELAS, the human-like attention generation task and classification task are jointly trained. Thus, we define a joint loss function in the training process upon the losses specified for different subtasks as follows:

$$J(\theta) = \sum(J_c(y, \hat{y}) + \lambda J_a(\text{HAM}, \text{MAM})). \tag{11}$$

where θ denotes, as a whole, the parameters used in our model, and λ is the hyper-parameter for striking a balance between document classification supervision and attention supervision. When only a few documents contain HAMs, the tunable parameter λ can be optimized to emphasize the small corresponding supervision signals, then both the classification and the human-like explanation goals can be achieved evenly.

During the training process, if there is no HAM for the input text, we only minimize Equation 10. When both HAMs and classification labels are available, we minimize Equation 11.

IV. Experiments

We evaluate our proposed method on four publicly available datasets that are compared against four state-of-the-art methods.

A. Datasets

The four datasets used in our experiments contain document labels for all instances while only a few of them have HAMs. All datasets contain roughly balanced examples between positive and negative classes. The proportions of positive examples are between 45% to 68%. It should be noted that our work can also be applied to multi-class datasets.

• **Yelp-HAT [11].** This dataset provides human attention maps for a collection of 1000 reviews from the Yelp dataset. Each review comes with a human attention map and a class label indicating whether the review is positive or negative. All characters are lowercase, punctuation is removed. Reviews are 50-75 words long. 70% of reviews are used for training with the remaining 30% for testing.

The dataset contains annotations from *multiple* humans for each of the reviews because each annotator may have different opinions on how indicative words are for review sentiments. To obtain reliable representations of human attention, we apply Consensus Attention Maps as being used in [11], by extracting HAMs from all annotators' agreement that are then used to evaluate a sentiment classification task.

• **N2C2.** N2C2 NLP Research datasets contain unstructured notes from the Research Patient Data Repository at Partners Healthcare[1]. From this clinical note repository, we use the 2014 challenge data, consisting of a set of medical documents that track the progression of heart disease in diabetic patients. Each clinical note is assigned to an expert in order to indicate the presence and progression of a disease (diabetes or heart disease), associated risk factors, and the time they were present in the patient's medical history.

In this dataset, we focus on predicting heart disease. For each patient in the dataset, if there is a clinical note with a heart

[1] https://n2c2.dbmi.hms.harvard.edu

disease annotation (indicated by CAD tag), we assign all notes belonging to this patient to the positive class. Patients with no heart disease mention are assigned to the negative class. Then we train a model that inputs every individual clinical note and predicts whether this note belongs to a heart-disease patient. N2C2 dataset contains 520 clinical notes in the training set and 511 clinical notes for the testing set. A series of notes from the same patient is assigned into either the training or testing set.

We use all heart disease-related words, as outlined by the annotation guidelines of 2014 Heart Disease Risk Factors Challenge of n2c2 NLP Research Data Sets[2], to create human attention maps. These include remarks of patients having heart disease (e.g., "coronary artery disease") or indirect mentions (e.g., "unstable angina," "PLAVIX" - a blood thinner used to prevent heart attack).

• **Movie Reviews [30].** Each review comes with a positive/ negative sentiment label and human annotation on word-level. Due to the length constraint of our model, we used the first 200 words as text input. Reviews in which the first 200 words are all labeled 0 are dropped. After preprocessing, there are 1,241 reviews in the training set and 320 reviews in the testing set.

• **Standard Sentiment Treebank (SST) [31].** This dataset contains 9,545 sentences in the training set and 2,310 sentences in the testing set. Each sentence comes with a binary classification label (positive or negative). The original data do not contain human attention annotation. We randomly selected 400 sentences from the dataset (200 from training split and 200 from testing split, positive/negative sentences ratio 1:1). Then we asked four researchers in our groups to annotate words that are indicative of the review sentiment in each sentence.

B. Compared Methods

We compare the proposed HELAS with the following methods:

• **Limited Supervised RA**. Rationale-Augmented (RA) model is proposed by [17]. The model has a hierarchical structure that first estimates the probability of each sentence in the input document is *rationale* (which means labeled as high attention in human attention map), in which the probability is determined by sentence representation without considering document representation. Then it produces a document-level representation by taking the sum of its constituent sentence representations weighed by these estimates. The document representation is further used to make document class prediction. The model is first trained with sentence-level prediction task then being trained with document-level prediction task. To compare against our method, we use the word-level binary attention (high attention or low attention) prediction task instead of sentence-level three rationale classes (positive rationale, negative rationale, or non-rationale) prediction task. Here, "Limited Supervised" means that the model is only trained with data that have both document classification labels and HAMs. Note that most of the data having only document classification labels are not used in the method.

• **Self-labeling RA**. This approach also utilizes the model framework proposed in [17]. Self-labeling was proposed by [32]. In our case, we repeat the word-level training phrase multiple runs. In each run, the model is trained on data with HAMs. After the model is trained, it generates pseudo HAMs for data without HAMs. Then instances with high confident predictions on HAMs are added to the training data. The procedure is repeated until training data without HAMs are exhausted. After that, the model is trained on the document classification task.

• **External Attention SCHA**. SCHA (Sequence classification with human attention) is proposed by [18]. Their model contains an attention mechanism that predicts attention weights for each word in the input document. This attention mechanism only takes word representations into account. The document representation is computed by taking the sum of its constituent word representations weighed by these estimates. It supervises the attention mechanism while training a classification task. This method assumes access to two unrelated datasets: One contains documents with classification labels, another one is a different set of documents contains HAMs (collected from human eye-tracking [33]). The training proceeds by flipping back and forth between two objectives depending on which dataset the training instance is drawn from. Note that in our implementation, this method does not have access to HAMs in the three classification datasets that we used in our experiments, it can only utilize HAMs from the human eye-tracking corpora.

• **Joint-learning SCHA.** This approach has the same training procedure as our proposed HELAS framework. The difference is that this approach used the model architecture proposed in [18] instead of our HELAS architecture.

For each method, we experiment with three text representation learners: LSTM, GRU, and BERT.

C. Metrics

The following two metrics are used for evaluation:

Behavioral Similarity [11]. To evaluate the explainable nature of each method, we use the *Behavioral Similarity* metric proposed in [11]. This metric measures the similarity between human and machine attention maps via the Area Under the ROC Curve:

$$B(\text{HAM}, \text{MAM}) = \frac{1}{|\mathcal{D}|} \sum_i \text{AUC}(\text{HAM}^i, \text{MAM}^i) \qquad (12)$$

where $|\mathcal{D}|$ is the number of documents in dataset $\mathcal{D}$. Behavioral similarity ranges between 0 and 1.

Accuracy. We use standard classification accuracy to measure the sequence classification performance.

D. Implementation Details

We implement the text representation learners as LSTM/GRU with 128-dimensional hidden states and BERT

[2] https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/

TABLE I: Performance of three text representation learners LSTM, GRU, and BERT on three tasks of Yelp-HAT Sentiment Classification, N2C2 Heart Disease Prediction, and Movie Reviews Sentiment Classification, with 2% of training data having HAMs. Metrics: (1) Behavioral similarity for human-like attention generation task, and (2) Accuracy for classification task.

Dataset	Methods	LSTM		GRU		BERT	
		Behave Sim.	Accuracy	Behave Sim.	Accuracy	Behave Sim.	Accuracy
Yelp-HAT	Limited Supervised RA	.62 ± .03	.56 ± .04	.72 ± .01	.57 ± .03	.40 ± .01	.77 ± .03
	Self-labeling RA	.75 ± .01	.88 ± .02	.79 ± .01	.89 ± .01	.59 ± .06	.94 ± .01
	External Attention SCHA	.75 ± .07	.65 ± .05	.82 ± .00	.89 ± .01	.76 ± .03	.95 ± .00
	Joint-learning SCHA	.57 ± .01	.67 ± .06	.57 ± .02	.89 ± .02	.57 ± .03	.95 ± .01
	HELAS (ours)	**.84** ± .00	**.92** ± .01	**.84** ± .00	**.92** ± .00	**.86** ± .01	**.96** ± .00
N2C2	Limited Supervised RA	.90 ± .01	.62 ± .05	.91 ± .00	.72 ± .01	.48 ± .05	.69 ± .01
	Self-labeling RA	.92 ± .00	.76 ± .00	.91 ± .01	.76 ± .00	.68 ± .06	.77 ± .00
	External Attention SCHA	.56 ± .02	.76 ± .00	.62 ± .02	.76 ± .00	.46 ± .05	.76 ± .00
	Joint-learning SCHA	.52 ± .06	.68 ± .00	.70 ± .07	.76 ± .00	.49 ± .05	.76 ± .00
	HELAS (ours)	**.93** ± .00	**.78** ± .00	**.92** ± .00	**.77** ± .00	**.73** ± .05	**.78** ± .01
Movie Reviews	Limited Supervised RA	.54 ± .01	.54 ± .00	.56 ± .03	.54 ± .00	.42 ± .01	.58 ± .03
	Self-labeling RA	.53 ± .02	.58 ± .04	.54 ± .02	.63 ± .06	.56 ± .03	.83 ± .02
	External Attention SCHA	.61 ± .02	.54 ± .00	.61 ± .01	.54 ± .00	.58 ± .01	.86 ± .01
	Joint-learning SCHA	.50 ± .02	.54 ± .00	.46 ± .01	.54 ± .00	.58 ± .02	.86 ± .00
	HELAS (ours)	**.69** ± .01	**.77** ± .00	**.69** ± .02	**.76** ± .01	**.80** ± .01	**.87** ± .01
SST	Limited Supervised RA	.81 ± .00	.66 ± .00	.88 ± .02	.68 ± .01	.82 ± .06	.78 ± .01
	Self-labeling RA	.84 ± .02	.71 ± .01	.86 ± .01	.72 ± .00	.96 ± .00	**.87** ± .00
	External Attention SCHA	.89 ± .00	.54 ± .00	.89 ± .00	.54 ± .00	.84 ± .04	**.87** ± .00
	Joint-learning SCHA	.50 ± .00	.54 ± .00	.50 ± .00	.54 ± .00	.47 ± .06	**.87** ± .00
	HELAS (ours)	**.91** ± .00	**.77** ± .00	**.91** ± .00	**.77** ± .00	**.97** ± .00	**.87** ± .00

[5]. The learning rates are 1e-3 and 2e-5 for LSTM/GRU and BERT, respectively. The LSTM/GRU model is trained for 40 epochs, while the BERT model is trained for 20 epochs. All three models are set the dropout rate at 0.2 and optimized using Adam [34]. We did a hyperparameter search for λ in the joint loss function. The best λ for LSTM model is 20, for GRU is 30, and for BERT is 4. All experiments are implemented on PyTorch [35] and run on a Tesla V100 GPU.

For Yelp-HAT dataset, we did a random train-test split every time. For N2C2, Movie Review, and SST datasets, we used the defined train-test splits every time. For Yelp-HAT, N2C2, and Movie Review datasets, the training data with and without HAMs are randomly assigned every time. We use the pre-trained BERT-base-uncased model from the "Transformers" library[3]. [36] For each experiment, we save the model with the highest accuracy during training and report the average evaluation results of each model from 5 replications that are initialized randomly. When we train a model with LSTM or GRU as the text representation learner, words are embedded using 100-dimensional GloVe [37] for Yelp-HAT and Movie Reviews. For N2C2 dataset, we use the pre-trained embeddings from BioMed [38]. When the text representation learner is BERT, we use the WordPiece embedding [29] provided by BERT model for all three datasets. All code and further training settings are publicly available[4].

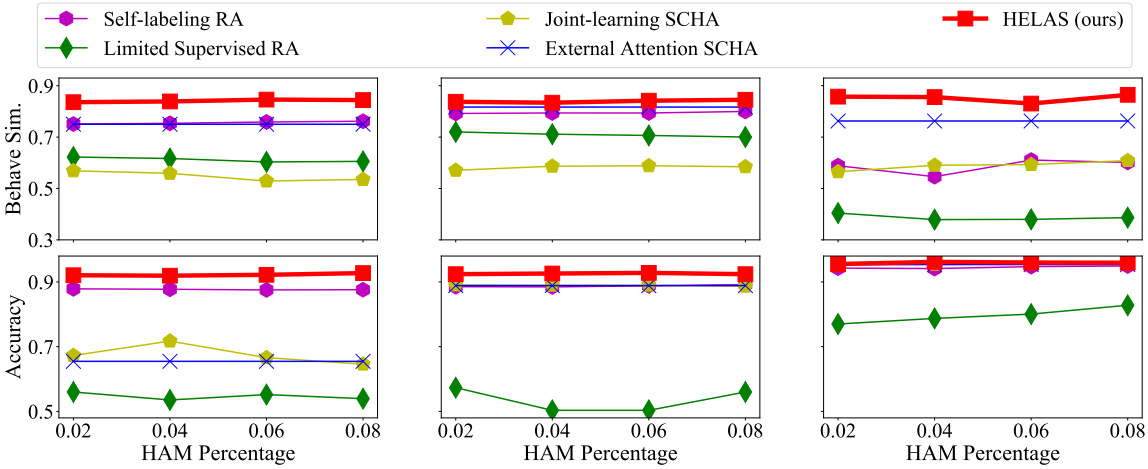
E. Experimental Results

1) Learning from Limited Labels: We first evaluate HELAS's capacity to learn from very limited levels of HAMs, specifically focusing on the case where only 2% of training data have HAMs. As always, the entire training dataset still has classification labels. In this experiment, we measure the behavioral similarity and the accuracy of all compared methods on the Yelp-HAT, N2C2, and Movie Reviews dataset, randomly down-sampling the HAM annotations to 2%. Our results are shown in Table I, where we first observe that our HELAS models achieve superior behavioral similarity and accuracy compared to all baseline models.

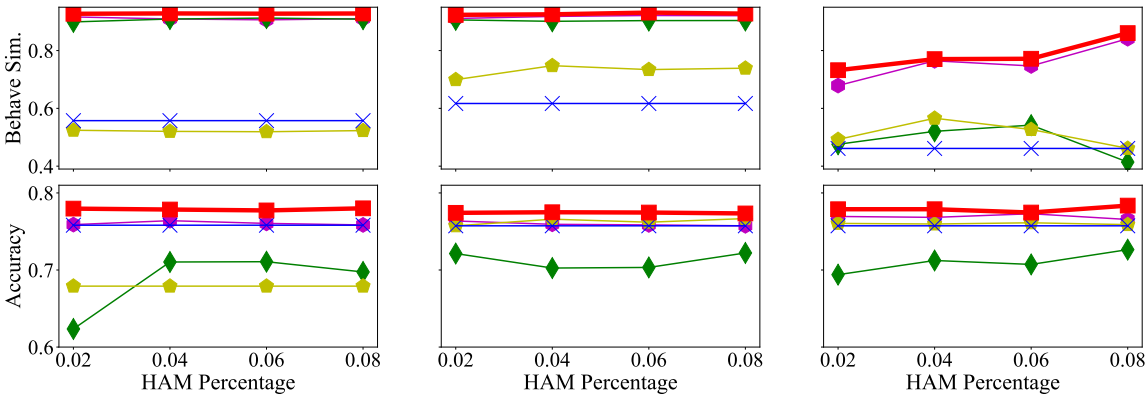
For the *Yelp-HAT sentiment classification task*, all HELAS models achieve significant gains in behavioral similarity (up to 9%) compared to the baseline models. HELAS with LSTM achieves the most substantial improvement in accuracy by 4%. Great gains in behavioral similarity indicate that the human-like attention learner in HELAS models can better mimic the relation between context and human-like attention even with a limited amount of word-level labels more. The HELAS models show improvement in the classification accuracy for all three core sequence algorithms, with HELAS-BERT achieving the least gain. This is likely because the HELAS-BERT model is

[3]https://github.com/huggingface/transformers

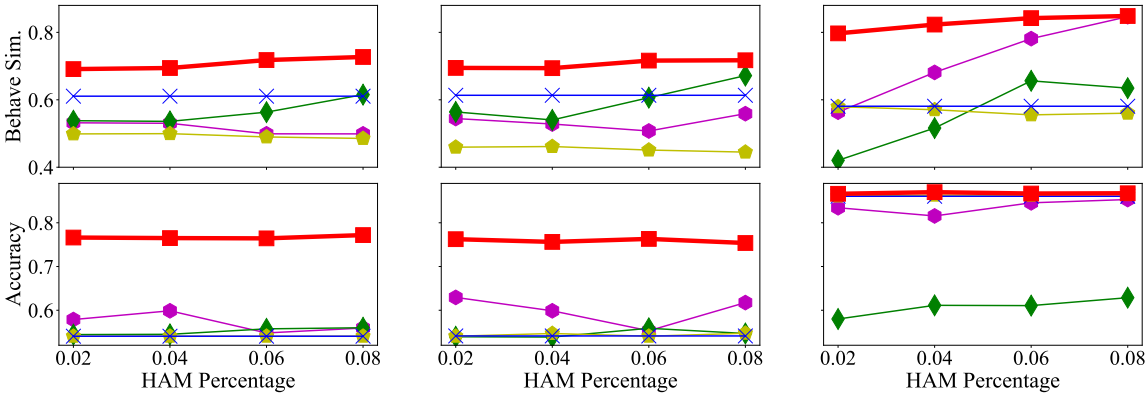
[4]https://github.com/zdy93/HELAS



(a) Yelp-HAT dataset with text representation learners: LSTM, GRU, and BERT shown on columns 1, 2 and 3, respectively.



(b) N2C2 dataset with text representation learners: LSTM, GRU, and BERT shown on columns 1, 2 and 3, respectively.



(c) Movie Reviews dataset with text representation learners: LSTM, GRU, and BERT shown on columns 1, 2 and 3, respectively.

Fig. 3: Compared performance on three datasets: Yelp, N2C2, and Movie Reviews. For each dataset, we experiment with three different text representation learner LSTM, GRU, and BERT. We vary the proportions of available HAMs in the training dataset as shown on the x-axis, ranging from 2%, 4%, 6%, and 8%. Metrics: (1) Behavioral similarity for human-like attention generation task, and (2) Accuracy for classification task are both plotted.

pre-trained on a large text corpus, whereas HELAS-LSTM and HELAS-GRU models are being trained from scratch on a small dataset. This causes these baseline BERT models to achieve an already high accuracy, which is challenging to improve upon.

For the *N2C2 heart disease prediction task* and *movie reviews sentiment classification task*, we observe similar trends as for the Yelp-HAT sentiment classification task. We observe the largest gains in the classification accuracy for HELAS-LSTM and HELAS-GRU models compared to HELAS-BERT over the baseline methods. Improvement in behavioral simi-

TABLE II: Performance of the proposed HELAS and its three variations on Yelp-HAT Sentiment Classification task, assuming only 2% of training data contain HAMs.

Methods	Behave Sim.	Accuracy
LSTM		
HELAS	**.84** ± .00	**.92** ± .01
HELAS-W	.83 ± .00	.90 ± .00
HELAS-S	.77 ± .03	.52 ± .00
HELAS-A	.58 ± .00	.85 ± .01
GRU		
HELAS	**.84** ± .00	**.92** ± .00
HELAS-W	.81 ± .00	.91 ± .01
HELAS-S	.78 ± .00	.55 ± .00
HELAS-A	.62 ± .01	.86 ± .00
BERT		
HELAS	**.86** ± .01	**.96** ± .00
HELAS-W	.85 ± .01	.95 ± .00
HELAS-S	.76 ± .03	.95 ± .01
HELAS-A	.74 ± .02	.94 ± .00

larity is significant (up to 22%) for all core algorithms.

For the *SST sentiment classification task*, the results show again that our method outperforms the other alternative methods on the behavior similarity. Both HELAS-LSTM and HELAS-GRU methods shown improvement in behavioral similarity by 2% and accuracy by around 5-6%. Most methods benefit from rich discriminative signals on this task and reach comparable performance when pairing with the BERT model.

Further results on other percentages of HAM availability are shown in Figure 3. Because we only labeled 400 sentences in the SST dataset, we did not conduct experiments on other percentages of HAM availability for the *SST sentiment classification task*. We observe that our HELAS models keep outperforming state-of-the-arts baselines across three tasks as the HAM proportion increases from 2% to 8%. Note that External Attention SCHA. utilizes an external source of HAMs and has no access to HAMs in classification task datasets, so its performance remains unchanged as HAM proportion increases.

2) Ablation Study: To test the performance gained by our customized Human-like Attention Learner and Contextualized Representation. We design three variations of our proposed methods, either context vector or attention function in each of which are different from HELAS model.

• **HELAS-W**: A variation of HELAS where we remove the document representation r from the Contextualized Representation c. This model may ignore information from some words with low attention.

• **HELAS-S**: A variation of HELAS where we remove the weighted sum of word representations $\sum_t \hat{\alpha}_t e_t$ from the Contextualized Representation c. This model cannot leverage word importance information from attention module.

• **HELAS-A**: A variation of HELAS where attention scores are defined as the similarity of word representation e_t with the vector representation r: $\hat{\alpha} = \text{sigmoid}(e_t^\top r)$.

Experimental results on Yelp-HAT dataset are presented in Table II. We test all three variations with three different text representation learners. HELAS outperforms both HELAS-W and HELAS-S, indicating that the combination of the document representation and the weighted sum of the word representation does indeed help to improve both the behavioral similarity and the classification accuracy. HELAS and HELAS-A performance prove that our specially-designed human-like attention learner can help our model better capture the reliance between context and human-like attention and utilize a limited amount of word-level labels more efficiently.

F. Case Study: Human-like Explanation

In this case study, our goal is to investigate whether human users prefer HELAS's MAMs as the best explanation for the text classification label comparing to MAMs generated by the other four compared methods. We conducted a user study involving 18 human participants, capable to read English text, with equal distribution in gender, and age ranging from 22 to 30 years, to evaluate MAMs on example documents. We randomly selected 15 examples from Yelp-HAT, N2C2, and movie review datasets (5 examples from each dataset). To keep results comparable, all methods use an LSTM as the text representation learner, and they are trained on data of which 2% have corresponding HAMs. Participants were instructed to choose the attention map that best provides clues about the classification label while maintaining the focus on only the most important words. Each participant was assigned 5 random example documents. For each example, we presented attention maps generated by all 5 methods; each with words with higher attention scores highlighted in a darker shade. We did not disclose the name of the method used for the attention generation. We randomly sort the choices to remove users' bias in the order of choices. The ground truth classification label is presented as a reference.

The results of this case study are illustrated in Table IV. We found that the number of times that HELAS's outputs were chosen by users is 57 out of 90 options, which is significantly higher than any other method. For each example, to conclude which method is more preferable, we use majority voting to aggregate all participants' selections. Out of 15 examples, the outputs from HELAS are selected as the best attention maps for 13 examples. It further confirms that a MAM generated by HELAS is more human-like than those by other models'. We note that this result is consistent with behavioral similarity results in Table I.

Table III visualizes the MAMs generated by HELAS-LSTM on examples taken from all four datasets. The HELAS-LSTM is trained on 2%-labeled data. The first example text is a review from the test set of the Yelp-HAT dataset. We observe that HELAS successfully assigns the highest attention weights to the most important words annotated by humans, which are underlined. Even at first glance, HELAS-generated MAMs

TABLE III: MAMs generated by HELAS. Test examples are from Yelp-HAT, N2C2, Movie Reviews, and SST dataset. Words are highlighted according to the attention scores. HAMs are shown in bold with underlines.

Dataset: **Yelp-HAT** Classification Label: **positive** review Food is **delicious,** the service is **fantastic,** and the rewards program is **ridiculously good.** My family eats here about once a week...
Dataset: **N2C2** Classification Label: the patient **had** heart disease **coronary artery disease status post non-ST elevation MI** , CHF with an EF of 35-40 % . PAST SURGICAL HISTORY : Includes a TAH , appendectomy , cataract surgery , ovarian cyst removal ...
Dataset: **Movie Reviews** Classification Label: is **positive** review Meet joe black is **a well acted romantic drama** which explores the meanings of life and love . William parrish (anthony hopkins) is a billionaire businessman on the brink of his 65th birthday ...
Dataset: **SST** Classification Label: is **positive** review **Laugh-out-loud lines** , **adorably ditsy but heartfelt performances** , and **sparkling** , **bittersweet dialogue** that cuts to the chase of the modern girl 's dilemma .

TABLE IV: Case study to evaluate human-like explanation. This table displays the number of times that a method's highlighted words were chosen by users as being the best explanation. For each document, user choices are aggregated by majority vote to determine the best method.

Methods	Total count of being selected	Majority vote on each example
Limited Supervised RA	4	0
Self-labeling RA	10	1
External Attention SCHA	15	1
Joint-learning SCHA	4	0
HELAS (ours)	**57**	**13**

provide clues about the classification. This review is labeled as positive since the food is delicious and the service is great at this restaurant.

The second example is from the N2C2 heart disease dataset. We again observe that HELAS focuses on the key information about the patient's condition and symptoms. The third and fourth examples are from the movie review and SST dataset. We notice that HELAS highlights the reviewers' compliments to the two movies, which can represent reviewers' positive attitudes towards these movies. In general, MAMs generated by each method can emphasize key information in the text related to the classification label.

V. Conclusion

In this paper, we define the open problem of explainable text classification with limited human attention supervision, with the aim to support the real-world setting in that human attention maps (HAMs) are often scarce. We propose the first solution to this problem, named HELAS: Human-like Explanation with Limited Attention Supervision. Our proposed method contains two key components: a human-like attention learner that successfully learns human-like attention weights conditioned on context information, and a carefully designed contextualized representation that considers the contribution from all words to classify the document into a final class. Our specially-designed joint loss function balances the supervision signals from both the *human-like attention generation* and *document classification* tasks simultaneously, despite them having drastically different numbers of labels across training instances.

Our evaluation studies on three real-world datasets demonstrate that HELAS outperforms state-of-the-art alternatives on both learning an accurate text classifier and generating human-like attention, even when as little as 2% of the data contain HAMs. This result is consistent across different text representation learners from LSTM, GRU, to BERT.

VI. Acknowledgment

We would like to thank the DSRG (Data Science Research Group) at Worcester Polytechnic Institute, and our reviewers for their feedback on the paper. We also thank the DAISY (DAta-driven Intelligent SYstem) Lab for their participation in our case study.

References

[1] R. K. Kaliyar, A. Goswami, P. Narang, and S. Sinha, "Fndnet–a deep convolutional neural network for fake news detection," *Cognitive Systems Research*, vol. 61, pp. 32–44, 2020.

[2] D. Zhang, J. Thadajarassiri, C. Sen, and E. Rundensteiner, "Time-aware transformer-based network for clinical notes series prediction," in *Machine Learning for Healthcare Conference*. PMLR, 2020, pp. 566–588.

[3] H. Zhang, S. Sun, Y. Hu, J. Liu, and Y. Guo, "Sentiment classification for chinese text based on interactive multitask learning," *IEEE Access*, vol. 8, pp. 129 626–129 635, 2020.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[6] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[7] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 1480–1489.

[8] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart, "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism," in *Advances in Neural Information Processing Systems*, 2016, pp. 3504–3512.

[9] Y. Sha and M. D. Wang, "Interpretable predictions of clinical outcomes with an attention-based recurrent neural network," in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, 2017, pp. 233–240.

[10] Y. Bao, S. Chang, M. Yu, and R. Barzilay, "Deriving machine attention from human rationales," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 1903–1913.

[11] C. Sen, T. Hartvigsen, B. Yin, X. Kong, and E. Rundensteiner, "Human attention maps for text classification: Do humans and neural networks focus on the same words?" in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[12] T. Qiao, J. Dong, and D. Xu, "Exploring human-like attention supervision in visual question answering," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[13] L. Chen, M. Zhai, and G. Mori, "Attending to distinctive moments: Weakly-supervised attention models for action localization in video," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 328–336.

[14] S. Liu, Y. Chen, K. Liu, and J. Zhao, "Exploiting argument information to improve event detection via supervised attention mechanisms," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2017, pp. 1789–1798.

[15] Y. Zhao, X. Jin, Y. Wang, and X. Cheng, "Document embedding enhanced event detection with hierarchical and supervised attention," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, vol. 2, 2018, pp. 414–419.

[16] M. Nguyen and T. Nguyen, "Who is killed by police: Introducing supervised attention for hierarchical lstms," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 2277–2287.

[17] Y. Zhang, I. Marshall, and B. C. Wallace, "Rationale-augmented convolutional neural networks for text classification," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, vol. 2016. NIH Public Access, 2016, p. 795.

[18] M. Barrett, J. Bingel, N. Hollenstein, M. Rei, and A. Søgaard, "Sequence classification with human attention," in *Proceedings of the 22nd Conference on Computational Natural Language Learning*, 2018, pp. 302–312.

[19] C. Liu, J. Mao, F. Sha, and A. Yuille, "Attention correctness in neural image captioning," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[20] I. Arous, L. Dolamic, J. Yang, A. Bhardwaj, G. Cuccu, and P. Cudré-Mauroux, "Marta: Leveraging human rationales for explainable text classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 7, 2021, pp. 5868–5876.

[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[22] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," *Syntax, Semantics and Structure in Statistical Translation*, p. 103, 2014.

[23] L. Liu, M. Utiyama, A. Finch, and E. Sumita, "Neural machine translation with supervised attention," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2016, pp. 3093–3102.

[24] S. Kuang, J. Li, A. Branco, W. Luo, and D. Xiong, "Attention focusing for neural machine translation by bridging source and target embeddings," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2018, pp. 1767–1776.

[25] S. Schuster and C. D. Manning, "Enhanced english universal dependencies: An improved representation for natural language understanding tasks," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016, pp. 2371–2378.

[26] T. Lei, R. Barzilay, and T. Jaakkola, "Rationalizing neural predictions," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 107–117.

[27] M. Yu, S. Chang, Y. Zhang, and T. Jaakkola, "Rethinking cooperative rationalization: Introspective extraction and complement control," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 4085–4094.

[28] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *Proceedings of International Conference on Learning Representations*, 2015.

[29] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[30] J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, and B. C. Wallace, "Eraser: A benchmark to evaluate rationalized nlp models," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[31] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. Manning, A. Ng, and C. Potts, "Parsing With Compositional Vector Grammars," in *EMNLP*, 2013.

[32] C. Rosenberg, M. Hebert, and H. Schneiderman, "Semi-supervised self-training of object detection models," 2005.

[33] N. Hollenstein, J. Rotsztejn, M. Troendle, A. Pedroni, C. Zhang, and N. Langer, "Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading," *Scientific data*, vol. 5, no. 1, pp. 1–13, 2018.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[35] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[36] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Huggingface's transformers: State-of-the-art natural language processing," *ArXiv*, vol. abs/1910.03771, 2019.

[37] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing*, 2014, pp. 1532–1543.

[38] S. Pyysalo, F. Ginter, H. Moen, T. Salakoski, and S. Ananiadou, "Distributional semantics resources for biomedical text processing," *Proceedings of LBM*, pp. 39–44, 2013.