Human Attention Maps for Text Classification: Do Humans and Neural Networks Focus on the Same Words?

Cansu Sen[1], Thomas Hartvigsen[2], Biao Yin[2], Xiangnan Kong[1,2], and Elke Rundensteiner[1,2]

[1]Computer Science Department , Worcester Polytechnic Institute
[2]Data Science Program, Worcester Polytechnic Institute
{*csen,twhartvigsen,byin,xkong,rundenst*}*@wpi.edu*

Abstract

Motivated by human attention, computational attention mechanisms have been designed to help neural networks adjust their focus on specific parts of the input data. While attention mechanisms are claimed to achieve interpretability, little is known about the actual relationships between machine and human attention. In this work, we conduct the first quantitative assessment of human versus computational attention mechanisms for the text classification task. To achieve this, we design and conduct a large-scale crowd-sourcing study to collect human attention maps that encode the parts of a text that humans focus on when conducting text classification. Based on this new resource of human attention dataset for text classification, YELP-HAT, collected on the publicly available YELP dataset, we perform a quantitative comparative analysis of machine attention maps created by deep learning models and human attention maps. Our analysis offers insights into the relationships between human versus machine attention maps along three dimensions: overlap in word selections, distribution over lexical categories, and context-dependency of sentiment polarity. Our findings open promising future research opportunities ranging from supervised attention to the design of human-centric attention-based explanations.

1 Introduction

Attention-based models have become the architectures of choice for a vast number of NLP tasks including, but not limited to, language modeling (Daniluk et al., 2017), machine translation (Bahdanau et al., 2015), document classification (Yang et al., 2016), and question answering (Kundu and Ng, 2018; Sukhbaatar et al., 2015). While attention mechanisms have been said to add interpretability since their introduction (Bahdanau et al., 2015), the investigation of whether this claim is correct has only just recently become a topic of high-interest (Mullenbach et al., 2018; Thorne et al., 2019; Serrano and Smith, 2019). If attention mechanisms indeed offer a more in-depth understanding of a model's inner-workings, application areas from model debugging to architecture selection would benefit greatly from profound insights into the internals of attention-based neural models.

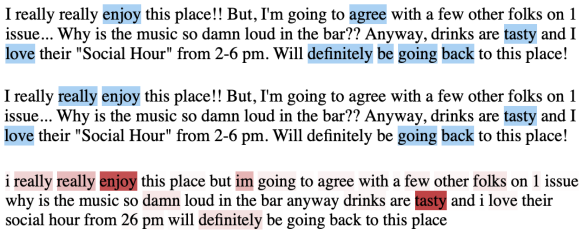

Figure 1: Examples of binary human attention (blue in top two texts) and continuous machine attention (red in bottom text).

Recently, Jain and Wallace (2019), Wiegreffe and Pinter (2019), and Serrano and Smith (2019) proposed three distinct approaches for evaluating the explainability of attention. Jain and Wallace (2019) base their work on the premise that explainable attention scores should be unique for a given prediction as well as consistent with other feature-importance measures. This prompts their conclusion that attention is not explanation. Based on similar experiments on alternative attention scores, Serrano and Smith (2019) conclude that attention does not necessarily correspond to the importance of inputs. In contrast, Wiegreffe and Pinter (2019) find that attention learns a meaningful relationship between input tokens and model predictions, which cannot be easily hacked adversarially.

While these works ask valuable questions, they embrace model-driven approaches for manipulating the attention weights and thereafter evaluate the

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4596–4608
July 5 - 10, 2020.

post-hoc explainability of the generated machine attention. In other words, they overlook the human factor in the evaluation process – which should be integral in assessing the plausibility of the generated explanations (Riedl, 2019).

In this work, we adopt a novel approach to attention explainability from a human-centered perspective and, in particular, investigate to what degree machine attention *mimics* human behavior. More precisely, we are interested in the following research question: *Do neural networks with attention mechanisms attend to the same parts of the text as humans?* To this end, we first collect a large dataset of human-attention maps and then compare the validated human attention with a variety of machine attention mechanisms for text classification.

Figure 1 displays examples of human and machine-generated attention for classifying a restaurant review's overall rating. Our goal is to quantify the similarity between human attention and machine-generated attention scores. Measuring this similarity is non-trivial and is not appropriately captured by an existing similarity metric (*e.g.*, Euclidean) between two vectors for the following reasons. A binary human attention vector does not solely denote which tokens are given higher importance but also implies information about the underlying grammatical structure and linguistic construction. For example, whether or not adjectives tend to be high-importance is encoded in the attention weights as well. Further, it is well known that human attention is itself subjective: given the same text and task, human annotators may not always agree on which words are important. That is, one single human's attention should rarely be regarded as the ground-truth for attention.

Given this objective, we use crowd-sourcing to collect a large set of human attention maps. We provide a detailed account of the iterative design process for our data collection study in §3. We design new metrics that quantify the similarity between machine and human attention from three perspectives (§4): *Behavioral similarity* measures the number of common words selected by human and machine discerning if neural networks with attention mechanisms attend to the same parts of the text as humans. Humans associate certain lexical categories (*e.g.*, adjectives) with a sentiment more heavily. *Lexical (grammatical) similarity* identifies if machine attention favors similar lexical categories with humans. A high lexical similarity shows that the attention mechanism learns similar language patterns with humans. *Context-dependency* quanitifies sentiment polarity of word selections.

We then employ these metrics to compare attention maps from a variety of attention-based Recurrent Neural Networks (RNN). We find that bi-Directional RNNs with additive attention demonstrate strong similarities to human attention for all three metrics. In contrast, uni-directional RNNs with attention differ from human attention significantly. Finally, as the text length increases, and with it, the prediction task becomes more difficult, both the accuracy of the models and similarity between human and machine decrease.

Our contributions are as follows:

- We conduct a large-scale collection of 15,000 human attention maps as a companion to the publicly-available Yelp Review dataset. Our collected Yelp-HAT (Human ATtention) dataset is publicly available as a valuable resource to the NLP community.
- We develop rich metrics for comparing human and machine attention maps for text. Our new metrics cover three complementary perspectives: *behavioral similarity*, *lexical similarity*, and *context-dependency*.
- We conduct the first in-depth assessment comparing human versus machine attention maps, with the latter generated by a variety of state-of-the-art soft and hard attention.
- We show that when used with bidirectional architectures, attention can be interpreted as human-like explanations for model predictions. However, as text length increases, machine attention resembles human attention less.

2 Preliminaries on Attention Maps

In this section, we define the concepts of Human Attention Map and Machine Attention Map.

Definition 2.1. Attention Map. An *Attention Map (AM)* is a vector where each entry in sequence is associated with a word in the corresponding position of the associated text. The value of the entry indicates the level of attention the corresponding word receives with respect to a classification task.

Definition 2.2. Human Attention Map. A Human Attention Map (HAM) is a binary attention

map produced by a human, where each entry with a set-bit indicates that the corresponding word receives high attention.

Definition 2.3. Machine Attention Map. A Machine Attention Map (MAM) is an attention map generated by a neural network model. If computed through *soft-attention*, a MAM corresponds to an AM of continuous values, that capture a probability distribution over the words. If computed through *hard-attention*, a MAM is a binary AM.

We now introduce the application of aggregation operators to coalesce HAMs by multiple annotators into aggregated HAMs.

Definition 2.4. Consensus Attention Map. If multiple HAMs exist for the same text, a Consensus Attention Map (CAM) is computed through a bitwise AND operation of the HAMs.

Definition 2.5. Super Attention Map. If multiple HAMs exist for the same text, a Super Attention Map (SAM) is computed by a bitwise OR operation of the HAMs.

3 Collection and Analysis of Human Attention Maps

3.1 HAM Collection by Crowd-sourcing

We collect human attention maps for the Yelp dataset[1] on the classification task of rating a review as positive or negative on Amazon Mechanical Turk. Participants are asked to complete two tasks: 1) Identify the sentiment of the review as positive, negative, or neither, and 2) Highlight the words that are indicative of the chosen sentiment. Our interface used for data collection is in Figure 2.

Preliminary investigation of the quality of human annotations. First, we conduct a series of data collection studies on two subsets of the Yelp dataset. Both subsets consist of 50 randomly-selected reviews from the *Restaurant* category. The first subset contains reviews with exactly 50 words, while the second contains reviews with exactly 100 words. For each review, human annotation is collected from two unique users.

We explore the quality of data we can collect on Mechanical Turk, as it encourages users to complete their tasks as quickly as possible since the number of completed tasks determines their income. This may lower the quality of collected

[1] https://www.yelp.com/dataset/challenge

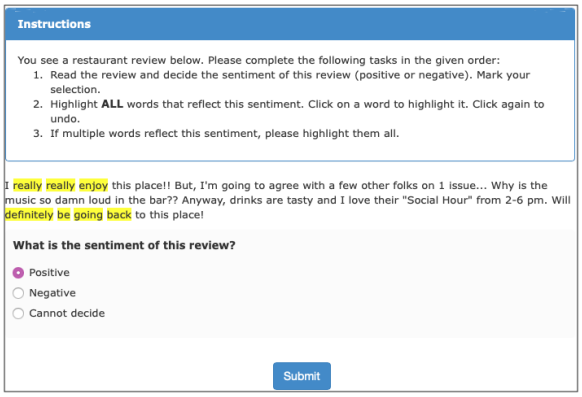

Figure 2: User interface we used for data collection on Amazon Mechanical Turk.

data since users may not select all relevant words, instead opting for the few most obvious ones, or they may choose words randomly.

Based on our preliminary investigations, we observe that both the average time users spend on the task (44 vs. 70 seconds) and the average number of words selected per review (9 vs. 13 words) increase as the number of words in the review increases from 50 to 100. This suggests that users do not choose words randomly; instead, they make an informed decision. We also visually examine the collected human attention maps and confirm that subjects make meaningful selections.

Pilot study assessing two design choices for data collection. Next, we design another pilot study to understand how humans perform the cognitive task of classifying a text and selecting the particular words that led to this decision. In this study, we ask eight participants to perform the same task while adhering to one of two strategies. The first strategy, the *read-first* design, involves reading the review first, deciding on the sentiment, then rereading the review, this time to highlight the relevant words. The second strategy, the *free-style* design, gives participants the freedom to choose the relevant words as they read the review to determine the sentiment. Each participant is asked to complete two tasks to experience both strategies. Half of the participants first work with the *read-first* design followed by the *free-style* design while the other half work in the reverse order. After completing the tasks, we ask the participants which strategy they find more natural in a post-task questionnaire.

Findings from the pilot study. Out of eight participants, half of them find it more useful reading the review first then deciding on the words whereas the other half indicated the opposite. We then evaluate

the collected data from three perspectives to decide which design is most suitable for our purposes.

We first examine the agreement between participants adhering to a particular strategy. This involves calculating the percentage of participants that mutually select the same phrase. We find that participant agreement is higher (73%) when the participants are forced to read the review before making any selections compared to using the free-style design (69%). Next, we investigate how similar the results are to the ground truth we defined for each review. The read-first design achieves better performance (3.30) compared to the free-style design (3.10). Our final criterion involves examining the amount of noise in the data (i.e., selections which deviate from the chosen sentiment). Only one review exhibits this situation where the review is clearly positive; however, it also contains a negative-opinion sentence. We observe that the read-first design reduces this cross-sentiment noise (1 vs. 0.5 scores).

Data collection protocol for the main study. Based on conclusions from the pilot studies, the *read-first* design is adopted to conduct the main data collection for $5,000$ reviews on Amazon Mechanical Turk. For this study, three different subjects annotated each review, resulting in a total of $15,000$ human attention maps. The resulting Yelp Human Attention Dataset (YELP-HAT) is publicly available [2] .

3.2 Analysis and Insights About HAMs

Factors that affect human accuracy. Some reviews contain a mixture of opinions, even though the reviewer felt strongly positive or negative about the restaurant. For example, consider the following review: *"Nothing to write home about, the chicken seems microwaved and the appetizers are meh. ... If your [sic] looking for a quick oriental fix I'd say go for it.. otherwise look elsewhere."* This review is labeled as *negative*, *positive*, and *neither*. The annotator who assigned it to the positive class selected the words "go for it" while the annotator who assigned it to the negative class selected the words "otherwise look elsewhere". This type of "mixed review" is the principal reason for discrepancies in classifications by the human annotators. The nature of crowd-sourcing also causes such inconsistencies as not all annotators provide reviews of equal quality.

Ambiguity in human attention. Intuitively, human attention is highly subjective. Some common patterns across annotators lead to differences in human annotations. A common behavior is to select *keywords* that indicate a sentiment. Another typical action is to select *entire sentences* if the sentence expresses an opinion.

Some reviews include subjective phrases that people interpret differently with regard to sentiment-polarity. For instance, "I come here often" can be construed as a favorable opinion. However, some people find it neutral. In some cases, an overwhelmingly-positive review incorporates a negative remark (or vice versa). In these cases, some people select all pieces of evidence of any sentiment, whereas others only choose words that indicate the prevailing sentiment.

4 Attention Map Similarity Framework

We quantify the similarity between HAMs and MAMs through our similarity framework that contains three new metrics as described in this section.

4.1 Overlap in Word Selections

For two attention mechanisms to be similar, they must put attention on the same parts of the text. Thus, we first define a metric for quantifying the overlap in the words selected by human annotators and by deep learning models.

Definition 4.1. Behavioral Similarity. Given a collection of attention maps $\text{HAM}_{\mathcal{D}}$ and $\text{MAM}_{\mathcal{D}}$ for a text dataset $\mathcal{D}$, behavioral similarity between human (H) and machine (M) corresponds to the average pair-wise similarity between each $(\text{HAM}_i, \text{MAM}_i)$ vector pair $\forall i \in \mathcal{D}$ as defined below:

$$\text{PairwiseSim}_i = AUC(\text{HAM}_i, \text{MAM}_i)$$

$$\text{BehavioralSim}(M, H) = \frac{1}{|\mathcal{D}|} \sum_i (\text{PairwiseSim}_i)$$

where $|\mathcal{D}|$ is the number of reviews in the dataset $\mathcal{D}$. Intuitively, this corresponds to adopting the human attention vector as binary ground truth. That is, it measures how similar the machine-generated continuous vector is to this ground truth. AUC is between 0 and 1 with .5 representing no similarity, and 1 the perfect similarity.

[2] http://davis.wpi.edu/dsrg/PROJECTS/YELPHAT/index.html

4.2 Distribution over Lexical Categories

Previous work has found that lexical indicators of sentiment are commonly associated with syntactic categories such as adjective, adverb, noun, and verb (Marimuthu and Devi, 2012). We define the following lexical similarity metric to test if human and machine adopt similar behaviors in terms favoring certain lexical categories.

Definition 4.2. Lexical Similarity. Given a collection of attention maps $\text{HAM}_{\mathcal{D}}$ and $\text{MAM}_{\mathcal{D}}$ for a text dataset $\mathcal{D}$, Lexical Similarity (LS) between human (H) and machine (M) over $\mathcal{D}$ is computed:

$$\text{LS}(M, H) = \text{corr}(\text{dist}(\text{words}_H), \text{dist}(\text{words}_M))$$

where words_H is a list of all selected words in all reviews of $\mathcal{D}$ by human, words_M is a list of all selected words in all reviews of $\mathcal{D}$ by machine, dist() is a function that computes the distribution of a word list over a tagset (e.g., nouns, verbs, etc.). After computing two distributions, the corr() function computes the correlation between them. In our experiments, we adopt Pearson Correlation. If MAM is continuous, selected words by M corresponds to k words with the highest attention scores, where k is the number of words selected by human for that text.

Using a *random attention* R as a baseline where the most important k words are selected randomly, we then compute an *Adjusted Lexical Similarity* which is between 0 and 1 as follows.

$$\text{AdjustedLS} = \frac{\text{LS}(M, H) - \text{LS}(R, H)}{1 - \text{LS}(R, H)}$$

4.3 Context-dependency of Sentimental Polarity

When deciding the sentiment of a review, human subjects may consider positive-sentiment words in a negative review and vice versa. To assess how context-dependant human and machine attentions are, we compute cross-sentiment selections rates.

Definition 4.3. Cross-sentiment selection rate (CSSR). Assume we have a collection of attention maps $\text{AM}_{\mathcal{D}}$ for a dataset $\mathcal{D}$, ground truth for overall sentiment Y for each review in $\mathcal{D}$ ($y_i \in \{0, 1\}$), and a list of positive words $\mathcal{P}$ and negative words $\mathcal{N}$ in the English language. CSSR denotes the ratio of selected words from the opposite sentiment.

$$p\_words = get\_words(HAM_{\mathcal{D}}, Y = 1)$$

$$n\_words = get\_words(HAM_{\mathcal{D}}, Y = 0)$$

$$CSSR_p = \frac{|p\_words \cap \mathcal{N}|}{|p\_words \cap \mathcal{P}|}$$

$$CSSR_n = \frac{|n\_words \cap \mathcal{P}|}{|n\_words \cap \mathcal{N}|}$$

get_words() function returns a list of attention-receiving words where $\text{HAM}_{ij} = 1, \forall i, j$ for the entire set of $\text{HAM}_{\mathcal{D}}$, for positive-sentiment reviews ($Y = 1$) and negative-sentiment reviews ($Y = 0$) separately. A list of words with positive and negative connotations, $\mathcal{P}$ and $\mathcal{N}$, are obtained from Hu and Liu (2004). CSSR_p (positive) and CSSR_n (negative) is then computed as the ratio of the number of cross-sentiment words over the number of same-sentiment words. A high CSSR means many words from the opposite sentiment are selected. This metric provides insights about how similar human and machine attentions are with regard to their context-dependant behaviour.

5 Is Machine Attention Similar to Human Attention?

5.1 Generating Machine Attention Maps

The Yelp dataset contains reviews and their rating scores between 0 and 5 (stars). This rating score corresponds to the ground truth for the review's overall sentiment. We create a binary classification task by assigning 1 and 2-star reviews to the negative class and 4 and 5-star reviews to the positive class. We omit 3-star reviews as they may not exhibit a clear sentiment. For training neural network models, we extract balanced subsets and split them into 80% training set, 10% validation set and 10% test sets. We then generate MAMs using the following machine learning models.

RNN with soft attention. Recurrent Neural Networks (RNN) enhanced with attention mechanisms have emerged as the state-of-the-art for NLP tasks (Bahdanau et al., 2015; Yang et al., 2016; Daniluk et al., 2017; Kundu and Ng, 2018). We implement the additive attention for many-to-one classification task as it is commonly used in the literature (Yang et al., 2016; Bahdanau et al., 2015) and paired it with both uni- and bi-directional RNN. In our implementation, we use LSTM memory cells.

	Accuracy		
	Yelp-50	Yelp-100	Yelp-200
Human	0.96	0.94	0.94
RNN	0.91 ± 0.006	0.90 ± 0.013	0.88 ± 0.01
biRNN	0.93 ± 0.008	0.91 ± 0.005	0.88 ± 0.02
Rationales	0.90 ± 0.004	0.85 ± 0.035	0.77 ± 0.015

Table 1: Test accuracy from three subsets of Yelp data.

Assuming that Γ is the recurrence function of LSTM and x_i is the embedded i-th word of T words in a review, we model our method as:

$$h_i = \Gamma(x_i, h_{i-1}), i \in [1, T] \quad (1)$$

$$u_i = \tanh(Wh_i + b) \quad (2)$$

$$\alpha_i = \frac{\exp(u_i^\top u)}{\sum_t \exp(u_i^\top u)} \quad (3)$$

Here $h_i, i \in [1, T]$ are hidden representations, W, b, and u are trainable parameters, and $\alpha_i, i \in [1, T]$ are the attention scores for each word x_i. A context vector c_i corresponds to the weighted average of the hidden representations of words with attention weights, denoted by:

$$c_i = \sum_j \alpha_j h_j \quad (4)$$

Through a softmax layer, context vector c_i is then used for further classifying the input sequence.

Rationale mechanism. An alternative approach, referred to as "rationale mechanism", can be seen as a type of hard attention (Lei et al., 2016; Bao et al., 2018). This model consists of two main parts that are jointly learned: a generator and an encoder. The generator specifies a distribution over the input text to select candidate rationales. The encoder is used to make predictions based on the rationales. The two components are integrated and regularized in the cost function with two hyper-parameters, selection lambda, and continuity lambda, for optimizing the representative selections. The selection lambda penalizes the number of words selected, while the continuity lambda encourages the continuity via minimizing the distances of the words chosen.

5.2 Behavioral Similarity Analysis

We conduct a set of controlled experiments with the length of the review changing across experiments. First, we generate MAMs for three subsets of the Yelp dataset: reviews containing 50 words (Yelp-50), 100 words (Yelp-100) and 200 words (Yelp-200). Neural network models with attention mechanisms are trained on each of these subsets. The corresponding test set accuracies for sentiment classification of human versus machine are shown in Table 1. Next, we acquire the HAMs collected for each test set. Since each review is annotated by three people, we have three sets of HAMs: HAM_1, HAM_2, and HAM_3. Consensus among the three, CAM and SAM, are computed as per Defs. 2.4 and 2.5. Then we measure the Behavioral Similarity between human and machine. The amount of overlap in the selected words are presented in Table 2.

We observe that accuracy and similarity both decrease as the review-length increases and the classification task becomes more difficult for both humans and machine learning models. We identify two reasons for this: First, when a review is long, the prevailing opinion is usually not obvious at first glance and may require more intensive reading and contemplating. Second, the reviewers are more likely to state conflicting facts and opinion in long reviews. This, in turn, creates distracting and hard-to-read text. Compared to unidirectional model, bidirectional RNN with attention consistently rates closer to human attention. This is most striking for the Yelp-50 subset. This can be explained with the fact that bidirectional RNNs possess information from both directions of the text similar to humans.

For all three subsets, Yelp-50, Yelp-100, and Yelp-200, behavioral similarity for Consensus Attention Map is higher than all three HAMs. This is an important result because it indicates that the words all annotators agreed to be important are selected by machine attention too, whereas more subjective selections do not always get high attention from machine, indicated by lower SAM similarity.

Finally, we compare similarity of these three sets of HAMs. Even though human-to-human similarity is usually higher than human-to-machine similarity (as expected), the numbers still far from being close to 1. This confirms the subjectivity of human attention. Also, note that human-to-human similarity decreases as the review length increases.

We observe that the performance of the rationale-based models degrades more sharply as the review-length increases. As our goal is to compare human attention with machine-generated attention for model interpretability, we optimize the model not only for accuracy but also for the number of selected rationales. We aim to generate roughly an

Yelp-50	$\text{HAM}_1, k=10$	$\text{HAM}_2, k=12$	$\text{HAM}_3, k=12$	$\text{CAM}, k=5$	$\text{SAM}, k=22$
HAM_2	0.73	-	-	-	-
HAM_3	0.74	0.75	-	-	-
RNN Attention	0.59 ± 0.021	0.59 ± 0.002	0.57 ± 0.012	0.59 ± 0.024	0.58 ± 0.021
Bi-RNN Attention	0.69 ± 0.004	0.70 ± 0.008	0.69 ± 0.007	0.79 ± 0.003	0.64 ± 0.008
Rationales	0.62 ± 0.014	0.62 ± 0.012	0.63 ± 0.015	0.68 ± 0.020	0.58 ± 0.010
Yelp-100	$\text{HAM}_1, k=15$	$\text{HAM}_2, k=16$	$\text{HAM}_3, k=16$	$\text{CAM}, k=6$	$\text{SAM}, k=30$
HAM_2	0.71	-	-	-	-
HAM_3	0.73	0.74	-	-	-
RNN Attention	0.57 ± 0.009	0.58 ± 0.011	0.59 ± 0.012	0.57 ± 0.010	0.58 ± 0.008
Bi-RNN Attention	0.65 ± 0.011	0.65 ± 0.021	0.66 ± 0.021	0.73 ± 0.031	0.62 ± 0.012
Rationales	0.55 ± 0.015	0.55 ± 0.005	0.55 ± 0.010	0.59 ± 0.015	0.54 ± 0.005
Yelp-200	$\text{HAM}_1, k=26$	$\text{HAM}_2, k=27$	$\text{HAM}_3, k=25$	$\text{CAM}, k=11$	$\text{SAM}, k=45$
HAM_2	0.70	-	-	-	-
HAM_3	0.69	0.71	-	-	-
RNN Attention	0.60 ± 0.011	0.60 ± 0.013	0.60 ± 0.014	0.60 ± 0.017	0.60 ± 0.011
Bi-RNN Attention	0.61 ± 0.015	0.61 ± 0.008	0.61 ± 0.018	0.63 ± 0.009	0.60 ± 0.008
Rationales	0.51 ± 0.013	0.52 ± 0.021	0.51 ± 0.018	0.52 ± 0.025	0.49 ± 0.019

Table 2: Behavioral similarity of human attention to machine on varying review length. k indicates the average number of words selected. (0.5:no similarity, 1.0:perfect similarity)

equal number of words selected by both human annotators and machine-generated rationales. Hence, we force the rationale-models to pick fewer words by tuning the selection lambda accordingly. This gives a comparative advantage to attention-based models against rationale-based models, as the rationale model is a hard-attention mechanism. In addition, rationales are better suited for sentence-level tasks as they encourage consecutive selection as opposed to the behavior of attention.

5.3 Lexical Similarity Analysis

Next, we analyze if humans and neural networks pay more attention to words from particular lexical categories using Adjusted Lexical Similarity score.

Lexical Similarity results, presented in Table 3, are consistent with Behavioral Similarity in that bidirectional model with attention is most similar to human (0.91 for Yelp-50 and 0.84 for Yelp-100). Rationales model follows bidirectional RNN, and unidirectional RNN is the least similar model to human. Overall, lexical similarity to human decreases for all models, as the reviews become longer.

Next, we inspect which lexical categories are selected more heavily by human and machine. For this, we provide relative frequency of lexical categories for human-selected words, machine-selected words (bi-RNN), and overall relative frequency of this tag within the dataset. Adjectives (Human:0.24 bi-RNN:0.23 Overall:0.02), comparative adjectives (Human:0.002 bi-RNN:0.001 Overall:0.0001), and nouns (Human:0.38 bi-RNN:0.37 Overall:0.09) are among the lexical categories that humans and bi-RNN models favor heavily. Similarly, personal pronouns are rarely selected by neither humans nor bi-RNN models (Human:0.005 bi-RNN:0.005 Overall:0.01).

5.4 Cross-sentiment Selection Rate Analysis

Finally, we compute CSSR scores, presented in Table 4, to evaluate the context-dependency of sentimental polarity for human and machine attentions. Our observations for Yelp-50 dataset are as follows. By human annotators, almost exclusively positive words are selected if the overall review sentiment is positive. For negative reviews, higher number of positive words are selected than negative words ($\text{CSSR}_p = 0.06, \text{CSSR}_n = 0.20$). Among the neural network models, the bidirectional RNN once more behaves most similar to human annotators with $\text{CSSR}_p = 0.04$ and $\text{CSSR}_n = 0.19$. RNN model's approach differs from that of human's and bi-RNN's. Even though the behaviour is similar for positive polarity ($\text{CSSR}_p = 0.06$), the opposite is true for negative polarity. In fact, positive words selected 2.28 times more than negative words in negative reviews, which is counter-intuitive. For the Rationales model, CSSR_p is 0.08 and CSSR_n

is 0.44. This indicates that Rationales model is more similar to human attention than RNN model with attention. We observe similar trends for the Yelp-100 and Yelp-200 datasets.

6 Related Work

A large body of work has been using attention mechanisms to attempt to bring 'interpretability' to model predictions (Choi et al., 2016; Sha and Wang, 2017; Yang et al., 2016). However, they only assess the produced attention maps qualitatively by visualizing a few hand-selected instances. Recently, researchers began to question the interpretability of attention. Jain and Wallace (2019) and Serrano and Smith (2019) argue that if alternative attention distributions exist that produce similar results to those obtained by the original model, then the original model's attention scores cannot be reliably used to explain the model's prediction. They empirically show that achieving such alternative distributions is possible. In contrast, Wiegreffe and Pinter (2019) find that attention learns a meaningful relationship between input tokens and model predictions which cannot be easily hacked adversarially.

Das et al. (2016) conducted the first quantitative assessment of computational attention mechanisms for the visual question answering (VQA) task. Similar to our work, they collect a human attention dataset, then measure the similarity of human and machine attention within the context of VQA. This VQA-HAT dataset now provides a fertile research vehicle for researchers in computer vision for studying the supervision of the attention mechanism (Liu et al., 2017a). The development of a similar dataset and an in-depth quantitative evaluation for text to advance NLP research is sorely lacking. In a concurrent and independent work, DeYoung et al. (2019) collects the ERASER dataset for human annotations of rationales. While ERASER includes multiple datasets for a number of NLP tasks with relatively small amounts of data for each, we focus on text classification and collect a large amount of data on a different corpus.

7 Discussion

Recent papers, including our work, take strides at answering the question if attention is interpretable. This is complicated by the fact that "interpretability" remains a not well-defined concept.

Attention adds transparency. Lipton (2018) defines *transparency* as overall human-understanding of a model, i.e., why a model makes its decisions. Under this definition, attention scores can be seen as partial transparency. That is, they provide a look into the inner workings of a model, in that they produce an easily-understandable weighting of hidden states (Wiegreffe and Pinter, 2019).

Attention is not faithful. Whether adversarial attention scores exist that result in the same predictions as the original attention scores helps us understand if attention is *faithful*. With their empirical analyses, Serrano and Smith (2019) and Jain and Wallace (2019) show that attention is not faithful.

Rationale models for human-like explanations. Riedl (2019) argues that *explanations* are post-hoc descriptions of how a system came to a given conclusion. This raises the question of what makes a good explanation of the behavior of a machine learning system. One line of research offers these explanations in the form of binary *rationales*, namely, explanations that plausibly justify a model's actions (Bao et al., 2018; Lei et al., 2016).

Our approach at attention as human-like explanations. In claiming *attention is explanation*, it is seen to mimic humans in rationalizing past actions. In our work, we approach interpretability from this human-centric perspective. We develop a systematic approach to either support or refute the hypothesis that attention corresponds to human-like explanations for model behavior. Based on our comparative analyses, we provide initial answers to this important question by finding insights into the similarities and dissimilarities of attention-based architectures to human attention.

Towards additional tasks beyond text classification. Confidently concluding whether attention mimics human requires tremendous efforts from many researchers with human data to be collected via a well-designed data collection methodology, both labor-intensive and costly task. In this work, we thus focus on one task, namely, sentiment classification, and collect HAM for this task and on a single dataset. We invite other researchers to continue this line of research by exploring other tasks (e.g., question answering).

Next steps in attention research. Our work opens promising future research opportunities. One is to supervise attention models explicitly. Attention mechanisms themselves are typically learned in an *unsupervised* manner. However, initial re-

	Yelp-50		Yelp-100		Yelp-200	
	Lexical Sim.	Adjusted LS	Lexical Sim.	Adjusted LS	Lexical Sim.	Adjusted LS
Random Attention	0.85 ± 0.006	-	0.84 ± 0.013	-	0.90 ± 0.010	-
RNN Attention	0.93 ± 0.015	0.54	0.91 ± 0.007	0.44	0.93 ± 0.005	0.37
Bi-RNN Attention	0.99 ± 0.005	0.91	0.98 ± 0.013	0.84	0.93 ± 0.003	0.36
Rationales	0.95 ± 0.012	0.66	0.93 ± 0.027	0.53	0.90 ± 0.002	0.05

Table 3: Lexical Similarity and Adjusted Lexical Similarity of human attention to machine on varying review length. (Adjusted LS 0:no similarity, 1:perfect similarity)

	$CSSR_p$	$CSSR_n$
Human	0.06	0.20
RNN Attention	0.06	2.28
Bi-RNN Attention	0.04	0.19
Rationales	0.08	0.44

Table 4: Cross-sentiment Selection Rates for positive and negative reviews for Yelp-50 dataset.

search offers compelling evidence for the success of supervised attention models (Chen et al., 2017; Liu et al., 2017b) in the computer vision area. Also, attention has the potential to be leveraged for both making predictions and concurrently producing human-centric explanations similar to rationale-based architectures.

8 Conclusion

To gain a deeper understanding of the relationships between human and attention-based neural network models, we conduct a large crowd-sourcing study to collect human attention maps for text classification. This human attention dataset represents a valuable community resource that we then leverage for quantifying similarities between human and attention-based neural network models using novel attention-map similarity metrics. Our research not only results in insights into significant similarities between bidirectional RNNs and human attention, but also opens the avenue for promising future research directions.

Acknowledgments

This research was supported by the U.S. Dept. of Education grant P200A150306, Worcester Polytechnic Institute through the Arvid Anderson Fellowship, and the National Science Foundation through grants IIS-1815866, IIS-1910880, IIS-1718310, and CNS -1852498. We thank Prof. Lane Harrison, WPI, for his advice and guidance on the design study for the data collection, and Prof. Jeanine Skorinko, WPI, for helpful discussion about the cognitive aspects of human attention.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *Proceedings of International Conference on Learning Representations*.

Yujia Bao, Shiyu Chang, Mo Yu, and Regina Barzilay. 2018. Deriving machine attention from human rationales. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1903–1913.

Lei Chen, Mengyao Zhai, and Greg Mori. 2017. Attending to distinctive moments: Weakly-supervised attention models for action localization in video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 328–336.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, pages 3504–3512.

Michał Daniluk, Tim Rocktäschel, Johannes Welbl, and Sebastian Riedel. 2017. Frustratingly short attention spans in neural language modeling. *ICLR*.

Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2016. Human attention in visual question answering: Do humans and deep networks look at the same regions? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 932–937.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2019. Eraser: A benchmark to evaluate rationalized nlp models. *arXiv preprint arXiv:1911.03429*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556.

Souvik Kundu and Hwee Tou Ng. 2018. A question-focused multi-factor attention network for question answering. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117.

Zachary C Lipton. 2018. The mythos of model interpretability. *Communications of the ACM*, 61(10):36–43.

Chenxi Liu, Junhua Mao, Fei Sha, and Alan Yuille. 2017a. Attention correctness in neural image captioning. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Shulin Liu, Yubo Chen, Kang Liu, and Jun Zhao. 2017b. Exploiting argument information to improve event detection via supervised attention mechanisms. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1789–1798.

K Marimuthu and Sobha Lalitha Devi. 2012. How human analyse lexical indicators of sentiments-a cognitive analysis using reaction-time. In *Proceedings of the 2nd Workshop on Sentiment Analysis where AI meets Psychology*, pages 81–90.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111.

Mark O Riedl. 2019. Human-centered artificial intelligence and machine learning. *arXiv preprint arXiv:1901.11184*.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.

Ying Sha and May D Wang. 2017. 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*, pages 233–240. ACM.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2019. Generating token-level explanations for natural language inference. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 963–969.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. 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)*, pages 11–20.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. 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*, pages 1480–1489.

A Appendix

A.1 Training Rationale-based models

For the Rationale Neural Prediction Framework, we use the Pytorch implementation[3] suggested by Lei et al. (2016). In this framework, the encoder is built as Convolutional Neural Network (CNN) and the generator is built as Gumbel Softmax with independent selectors. The following hyper-parameters of CNN are used as pointed out by (Lei et al., 2016): 200 hidden dimensions, 0.1 dropout rate, 2 hidden layers, 128 batch size, 64 epochs, 0.0003 initial learning rate.

We conducted an extensive parameter search to find the optimum values for the two key hyper-parameters of the rationale model, selection-lambda, and continuity-lambda, which regularize the number and the continuity of words selected during the optimization process. For the selection lambda, we experimented with values 1, 1e-1, 1e-2, 1e-3, 1e-4, 1e-5, 1e-6, 1e-7, 1e-8, 1e-9, and 0. For the continuity lambda, we experimented with values 0 and two times of selection lambda. We observe that the performance of the rationale-based model is extremely sensitive to its hyper-parameters.

One conflicting interest with the rationale-based models is that the more words the model selects, the accuracy becomes higher. As our goal is to compare human attention with machine-generated attention for model interpretability, we optimize the model not only for accuracy but also for the number of selected rationales. We aim to generate roughly an equal number of words selected by both human annotators and machine-generated rationales.

A.2 Training Attention-based models

We used the following hyper-parameters to RNN-based models. 100 hidden dimensions, 100 attention size, 0.2 dropout rate, 128 batch size, 64 epochs, 0.0001 initial learning rate.

A.3 Additional Analysis Results

An example visualization of the attention maps annotated by human annotators and machine learning models is provided in Figure 4. The agreement between human annotators and all machine learning models can be considered high in this example, as there are many mutual selections.

[3] https://github.com/yala/text_nn

Great place to go to have some drinks and food with a group of friends before heading out for the night. Service is mediocre and most of the menu is on the wall which was frustrating but the menu is very diverse. Really enjoyed the wagu carpaccio and the butterfish. Toro and uni were very fresh although the portions were small. One of the coolest things about this place is the fact that the pitchers of beer have a ice chamber in the center to keep your beer colder for longer! Genius! The wait can be long so come early.

Great place to go to have some drinks and food with a group of friends before heading out for the night. Service is mediocre and most of the menu is on the wall which was frustrating but the menu is very diverse. Really enjoyed the wagu carpaccio and the butterfish. Toro and uni were very fresh although the portions were small. One of the coolest things about this place is the fact that the pitchers of beer have a ice chamber in the center to keep your beer colder for longer! Genius! The wait can be long so come early.

Figure 3: Human attention is highly subjective. Some annotators tend to select only a few words, whereas others choose entire sentences.

Another example is provided in Figure 3, demonstrating the attention maps provided by two different annotators for the same review. This is an extreme example of the subjectivity of human attention. The first annotator only highlights individual words with the strongest cues of sentiment, whereas the second annotator sometimes selects entire sentences when they indicate a sentiment.

Table 5 shows the distribution of selected words over lexical categories for Human (CAM), Machine (bi-RNN), and the entire corpus for the Yelp-50 subset. Any divergence in the Human and Machine columns from the Corpus column indicates a tendency of selection for a lexical category. For example, adjectives are selected very heavily by both Human and Machine, even though they only make 0.02 of all words in the dataset.

Lexical Category	Human	Machine(bi-RNN)	Corpus
Coordinating conjunction	0.0000	0.0098	0.0147
Cardinal number	0.0098	0.0077	0.0043
Determiner	0.0112	0.0168	0.0312
Existentialthere	0.0000	0.0000	0.0000
Foreign word	0.0000	0.0000	0.0000
Preposition or subordinating conjunction	0.0266	0.0084	0.0298
Adjective	0.2374	0.2269	0.0201
Adjective, comparative	0.0021	0.0014	0.0002
Adjective, superlative	0.0252	0.0287	0.0016
List item marker	0.0000	0.0000	0.0000
Modal	0.0035	0.0000	0.0030
Noun, singular or mass	0.3838	0.3711	0.0950
Noun, plural	0.0000	0.0000	0.0000
Proper noun, singular	0.0000	0.0000	0.0000
Proper noun, plural	0.0413	0.0665	0.0154
Predeterminer	0.0000	0.0000	0.0000
Possessive ending	0.0000	0.0000	0.0000
Personal pronoun	0.0056	0.0049	0.0141
Possessive pronoun	0.0035	0.0028	0.0067
Adverb	0.1296	0.0931	0.0277
Adverb, comparative	0.0070	0.0000	0.0014
Adverb, superlative	0.0000	0.0000	0.0000
Particle	0.0000	0.0000	0.0000
Symbol	0.0000	0.0000	0.0000
to	0.0035	0.0007	0.0077
Interjection	0.0000	0.0000	0.0000
Verb, base form	0.0196	0.0028	0.0098
Verb, past tense	0.0070	0.0609	0.0148
Verb, gerund or present participle	0.0357	0.0462	0.0053
Verb, past participle	0.0455	0.0455	0.0083
Verb, non-3rd person singular present	0.0000	0.0028	0.0023
Verb, 3rd person singular present	0.0007	0.0021	0.0065
Wh-determiner	0.0000	0.0000	0.0005
Wh-pronoun	0.0007	0.0000	0.0005
Possessive wh-pronoun	0.0000	0.0000	0.0000
Wh-adverb	0.0007	0.0007	0.0012

Table 5: Distribution over lexical categories for human-selected words, machine-selected words, and the entire corpus.

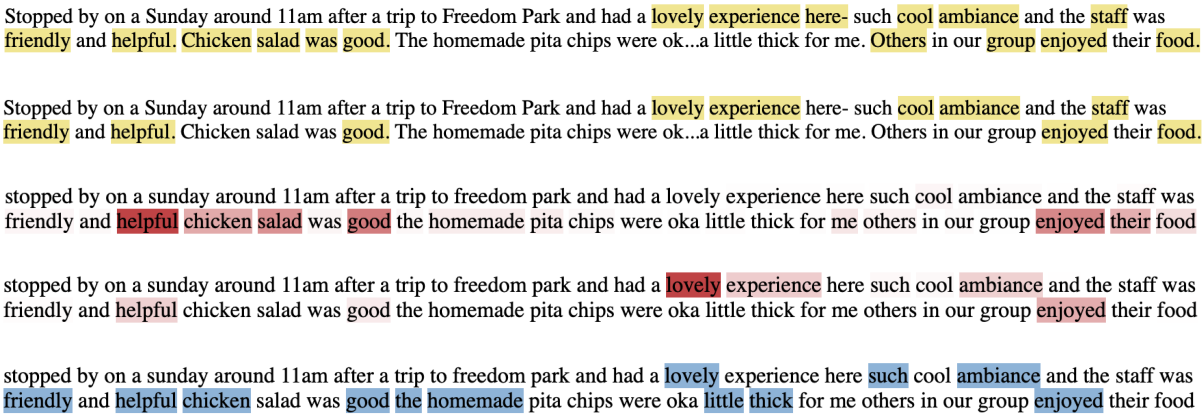

Figure 4: Visualizations of attention maps by human annotators and machine learning models. From top to bottom: first human annotator, second human annotator, RNN, bi-RNN, Rationales.