Why and when should you pool? Analyzing Pooling in Recurrent Architectures

Pratyush Maini[†], **Keshav Kolluru**[†], **Danish Pruthi**[‡], **Mausam**[†]
[†]Indian Institute of Technology, Delhi, India
[‡]Carnegie Mellon University, Pittsburgh, USA
{pratyush.maini, keshav.kolluru}@gmail.com,
ddanish@cs.cmu.edu, mausam@cse.iitd.ac.in

Abstract

Pooling-based recurrent neural architectures consistently outperform their counterparts without pooling on sequence classification tasks. However, the reasons for their enhanced performance are largely unexamined. In this work, we explore three commonly used pooling techniques (mean-pooling, max-pooling, and attention[1]), and propose *max-attention*, a novel variant that captures interactions among predictive tokens in a sentence. Using novel experiments, we demonstrate that pooling architectures substantially differ from their non-pooling equivalents in their learning ability and positional biases: (i) pooling facilitates better gradient flow than BiLSTMs in initial training epochs, and (ii) BiLSTMs are biased towards tokens at the beginning and end of the input, whereas pooling alleviates this bias. Consequently, we find that pooling yields large gains in low resource scenarios, and instances when salient words lie towards the middle of the input. Across several text classification tasks, we find max-attention to frequently outperform other pooling techniques.[2]

1 Introduction

Pooling mechanisms are ubiquitous components in Recurrent Neural Networks (RNNs) used for natural language tasks. Pooling operations consolidate hidden representations from RNNs into a single sentence representation. Various pooling techniques, like mean-pooling, max-pooling, and attention, have been shown to improve the performance of RNNs on text classification tasks (Lai et al., 2015; Conneau et al., 2017). Despite widespread adoption, precisely how and when pooling benefits the models is largely under-explored.

In this work, we perform an in-depth analysis comparing popular pooling methods, and proposed max-attention, with standard BiLSTMs for several text classification tasks. We identify two key factors that explain the benefits of pooling techniques: learnability, and positional invariance.

First, we analyze the flow of gradients for different classification tasks to assess the learning ability of BiLSTMs (§ 5). We observe that the gradients corresponding to hidden representations in the middle of the sequence vanish during the initial epochs. On training for more examples, these gradients slowly recover, suggesting that the gates of standard BiLSTMs require many examples to learn. In contrast, we find the gradient norms in pooling-based architectures to be free from this problem. Pooling enables a fraction of the gradients to directly reach any hidden state instead of having to backpropagate through a long series of recurrent cells. Thus we hypothesize, and subsequently confirm, that pooling is particularly beneficial for tasks with long input sequences.

Second, we explore the positional biases of BiLSTMs, with and without pooling (§ 6). Across several classification tasks, and various novel experimental setups, we expose that BiLSTMs are less responsive to tokens towards the middle of the sequence, when compared to tokens at the beginning or the end of the sequence. However, we find that this bias is largely absent in pooling-based architectures, indicating their ability to respond to salient tokens regardless of their position.

Third, we propose max-attention, a novel pooling technique, which combines the advantages of max-pooling and attention (§ 3.2). Max-attention uses the max-pooled representation as its query vector to compute the attention weights for each hidden state. Max-pooled representations are extensively used in the literature to capture prominent tokens (or objects) in a sentence (or an im-

[1]Attention aggregates representations via a weighted sum, thus we consider it under the umbrella of pooling in this paper.

[2]Code and data is made available at https://github.com/dair-iitd/PoolingAnalysis.

age) (Zhang and Wallace, 2015; Boureau et al., 2010b). Therefore, using them as a query vector effectively captures interactions among salient portions in the input. Max-attention is simple to use, and yields performance gains over other pooling methods on several classification setups.

2 Related Work

Pooling: A wide body of work compares the performance of different pooling techniques in object recognition tasks (Boureau et al., 2010a,b, 2011) and finds max-pooling to generally outperform mean-pooling. However, pooling in natural language tasks is relatively understudied. For some text classification tasks, pooled recurrent architectures (Lai et al., 2015; Zhang and Wallace, 2015; Johnson and Zhang, 2016; Jacovi et al., 2018; Yang et al., 2016a), outperform CNNs and BiLSTMs. Additionally, for textual entailment tasks, Conneau et al. (2017) find that max-pooled representations better capture salient words in a sentence. Our work extends the analysis and examines several pooling techniques, including attention, for BiLSTMs applied to natural language tasks. While past approaches assess the ability of pooling in capturing linguistic phenomena, to the best of our knowledge, we are the first to systematically study the training advantages of various pooling techniques.

Attention: First proposed as a way to align target tokens to the source tokens in translation (Bahdanau et al., 2014), the core idea behind attention—learning a weighted sum of the hidden states—has been widely adopted. As attention aggregates hidden representations, we consider it under the umbrella of pooling. Recently, Pruthi et al. (2020) conjecture that attention offers benefits during training; our work explains, and provides empirical evidence to support the speculation.

Gradient Propagation: Vanilla RNNs are known to suffer from the problem of vanishing and exploding gradients (Hochreiter, 1991; Bengio et al., 1994). In response, Hochreiter and Schmidhuber (1997) invented LSTMs, which provide a direct connection passage through all the cells in order to remember new inputs without forgetting prior history. However, recent work suggests that LSTMs do not solve this problem completely (Arjovsky et al., 2015; Chandar et al., 2019). Our work quantitatively investigates this phenomenon, exposing scenarios where the effect is pronounced, and demonstrating how pooling techniques mitigate the problem, leading to better sample efficiency, and generalization.

3 Methods

3.1 Background and Notation

Let $s = \{x_1, x_2, \ldots, x_n\}$ be an input sentence, where x_t is a representation of the input word at position t. A recurrent neural network such as an LSTM produces a hidden state h_t, and a cell state c_t for each input word x_t, where $h_t, c_t = \phi(h_{t-1}, c_{t-1}, x_t)$. Standard BiLSTMs concatenate the first hidden state of the backward LSTM, and the last hidden state of the forward LSTM for the final sentence representation: $s_{\text{emb}} = [\overrightarrow{h_n}, \overleftarrow{h_1}]$. The sentence embedding (s_{emb}) is further fed to a downstream text classifier. For training BiLSTMs, multiple works have emphasized the importance of initializing the bias for forget gates to a high value (between 1-2) to prevent the model from forgetting information before it learns what to forget (Gers et al., 2000; van der Westhuizen and Lasenby, 2018). Hence, in our analysis, we experiment with both a high and low value of bias for the forget gate. For the non-pooled BiLSTM, we initialize the forget gate bias to 1, unless specified. For brevity, from hereon we would use h_t to mean $[\overrightarrow{h_t}, \overleftarrow{h_t}]$. Below, we formally discuss popular pooling techniques:

Max-pooling: For a max-pooled BiLSTM (MAXPOOL), the sentence embedding s_{emb}, is:

$$s^i_{\text{emb}} = \max_{t \in (1,n)} (h^i_t)$$

where h^i_t represents the i^{th} dimension of the hidden state corresponding to the word at position t. This implies that while backpropagating the loss, we find a direct pathway to the t^{th} hidden state as:

$$\frac{\partial s^i_{\text{emb}}}{\partial h^i_t} = \begin{cases} 1, & \text{if } t = \text{argmax}_{\text{t} \in (1,\text{n})}\, \text{h}^\text{i}_\text{t} \\ \frac{\partial h^i_k}{\partial h^i_t}, & \text{if } k = \text{argmax}_{\text{t} \in (1,\text{n})}\, \text{h}^\text{i}_\text{t}, \text{k} \neq \text{t} \end{cases}$$

Similarly, in **mean-pooling** (MEANPOOL), s_{emb} is an average over all the hidden states.[3]

Attention: Attention (ATT) works by calculating a non-negative weight for each hidden state that together sum to 1. Hidden representations are then

[3]Refer to Appendix A.3 for the mathematical formulation.

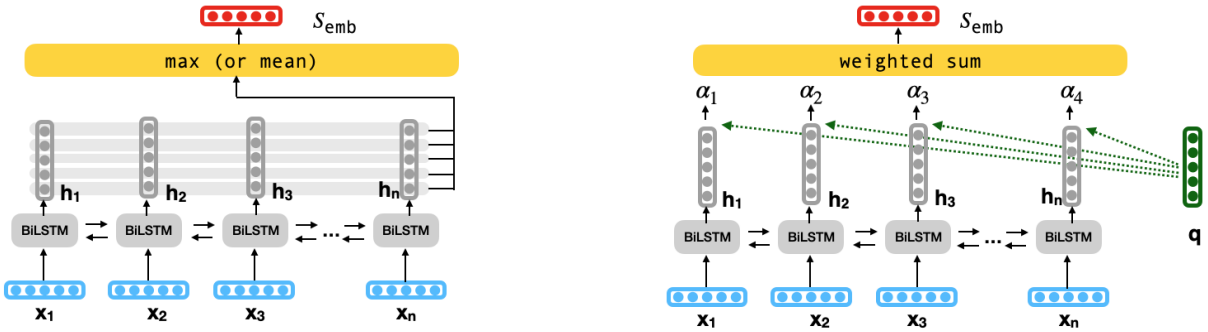


Figure 1: A pictorial overview of the pooling techniques. Left: element-wise mean and max pooling operations aggregate hidden representations. Right: attention scores (α) are computed using the similarity between hidden representations (h) and query vector (q), which are subsequently used to weight hidden representations. Our proposed max-attention uses the sentence embedding from max-pooling as a query to attend over hidden states.

multiplied with these weights and summed, resulting in a fixed-length vector (Bahdanau et al., 2014; Luong et al., 2015):

$$\alpha_t = \frac{\exp(h_t^\top q)}{\sum_{j=1}^{n} \exp(h_j^\top q)}; \quad s_{\text{emb}} = \sum_{t=1}^{n} \alpha_t h_t$$

where q is a learnable query vector. Several variations like hierarchical attention (Yang et al., 2016b), self-attention (Madasu and Rao, 2019) have been proposed for text classification. However, the above formulation (referred in literature as "Luong attention") is most widely used in text classification tasks (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019; Pruthi et al., 2020).

3.2 Max-attention

We introduce a novel pooling variant called max-attention (MAXATT) to capture inter-word dependencies. It uses the max-pooled hidden representation as the query vector for attention. Formally:

$$q^i = \max_{t \in (1,n)} (h_t^i); \qquad \hat{h_t} = h_t / \|h_t\|$$

$$\alpha_t = \frac{\exp(\hat{h_t}^\top q)}{\sum_{j=1}^{n} \exp(\hat{h_j}^\top q)}; \quad s_{\text{emb}} = \sum_{t=1}^{n} \alpha_t h_t$$

It is worth noting that the learnable query vector in Luong attention is the same for the entire corpus, whereas in max-attention each sentence has a unique locally-informed query. Previous literature extensively uses max-pooling to capture the prominent tokens (or objects) in a sentence (or image). Hence, using max-pooled representation as a query for attention allows for a second round of aggregation among important hidden representations.

3.3 Transformers

We briefly experiment with transformer architectures (Vaswani et al., 2017; Devlin et al., 2018), and observe that purely attention-based architectures perform poorly on text-classification without significant pre-training. Further, the memory footprint for transformers is $O(n^2)$ vs $O(n)$ for LSTMs. Thus, for long examples used in some of our experiments ($\sim$ 4000 words), XL-Net (Yang et al., 2019) runs out of memory even for a batch size of 1 on a 32GB GPU.

We observe that *CLS*-based text classification with pretrained transformers (such as RoBERTa (Liu et al., 2019)) results in near state-of-art performance. Alternate classification techniques using pooled feature representations result in a marginal difference in performance ($\sim 0.2\%$ on IMDB sentiment analysis). Pooling does not benefit transformers as they do not suffer from vanishing gradients and positional biases which pooling helps to mitigate in LSTMs (§ 5,§ 6). Therefore, we limit the scope of this work to recurrent architectures.

4 Datasets & Experimental Setup

We experiment with four different text classification tasks: (1) The **IMDb** dataset (Maas et al., 2011) contains movie reviews and their associated sentiment label; (2) **Yahoo! Answers** (Zhang et al., 2015) dataset comprises 1.4 million question and answer pairs, spread across 10 topics, where the task is to predict the topic of the answer, using the answer text; (3) **Amazon** reviews (Ni et al., 2019) contain product reviews from the Amazon website, filtered by their category. We construct a 20-class classification task using these reviews[4]; (4) **Yelp**

[4]Appendix B.1 contains further details about the dataset.

Reviews (Zhang et al., 2015) is another sentiment polarity classification task.

For these datasets, we only use the text and labels, ignoring any auxiliary information (like title or location). We select subsets of the datasets with sequences having greater than 100 words to better understand the impact of vanishing gradients and positional bias in recurrent architectures. A summary of statistics is presented in Table 1.

Dataset	Classes	Avg. Length	Max Length	Train Size	Test Size
IMDb	2	240.4	2470	20K	9.8K
Yahoo! Answers	10	206.2	998	25K	4.8K
Amazon Reviews	20	185.6	500	25K	12.5K
Yelp Reviews	2	202.4	1000	25K	9.5K

Table 1: Corpus statistics for classification tasks.

In all the experiments, we use a single-layered BiLSTM with hidden dimension size of 256 and embedding dimension size of 100 (initialized with GloVe vectors (Pennington et al., 2014) trained on a 6 billion word corpus). The sentence embeddings generated by the BiLSTM are passed to a final classification layer to obtain per-class probability distributions. We train our models using Adam optimizer (Kingma and Ba, 2014), with a learning rate of 2×10^{-3}. The batch size is set to 32 for all the experiments. We train for 20 epochs and select the model with the best validation accuracy. All experiments are repeated over 5 random seeds using a single GPU (Tesla K40).[5]

5 Gradient Propagation

In this section, we study the flow of gradients in different architectures and training regimes. Pooling techniques used in conjunction with BiLSTMs provide a direct gradient pathway to intermediate hidden states. However for BiLSTMs without pooling, it is crucial that the parameters for the input, output, and forget gates are appropriately learned so that the loss backpropagates across long input sequences, without the gradients vanishing.

Experimental Setup: In order to quantify the extent to which the gradients vanish across different word positions, we compute the gradient of the loss function w.r.t the hidden state at every word position t, and study their ℓ_2 norm ($\|\frac{\partial L}{\partial h_t}\|$). To aggregate the gradients across multiple training examples (of different lengths), we linearly interpolate the distribution of gradient values for each example to a fixed length between 1 and 100. The gradient values at each (normalized) position are averaged across all the training examples. We plot these values (on a log scale) after training on the first 500 IMDb reviews to study the effect of gradient vanishing at the beginning of training (Figure 2a).

To understand how the distribution of gradients (across word positions) changes with the number of training batches, we compute the ratio of the gradient norm corresponding to the word at the middle and word at the end: $\|\frac{\partial L}{\partial h_{\text{mid}}}\| \;/\; \|\frac{\partial L}{\partial h_{\text{end}}}\|$.[6] We call this the *vanishing ratio* and use it as a measure to quantify the extent of vanishing (where lower values indicate severe vanishing). Each training batch on the x-axis in Figures 2b, 2c corresponds to 64 training examples.

Results It is evident from Figure 2a that the gradients vanish significantly for BiLSTM, with $\|\frac{\partial L}{\partial h_t}\|$ falling to the order of 10^{-6} as we approach the middle positions in the sequence. This effect is even more pronounced for the case of $\text{BiLSTM}_{\text{LowF}}$, which uses the Xavier initialization (Glorot and Bengio, 2010) for the bias of the forget-gate. The plot suggests that specific initialization of the gates with best practices (such as setting the bias of forget-gate to a high value) helps to reduce the extent of the issue, but the problem still persists. In contrast, none of the pooling techniques face this issue, resulting in an almost straight line.

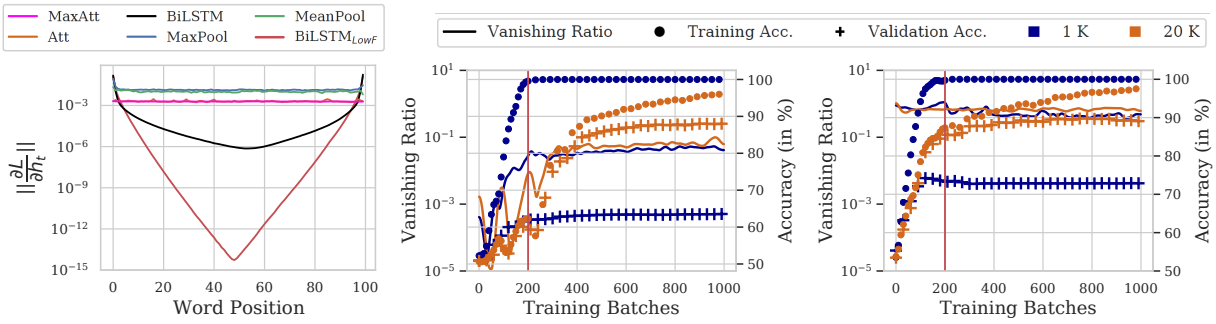
Additionally, from Figure 2b we note that the problem of vanishing gradients is more pronounced at the beginning of training, when the gates are still untrained. The problem continues to persist, albeit to a lesser degree, until later in the training process. This specifically limits the performance of BiLSTM in resource-constrained settings, with fewer training examples. For instance, in the 1K training data setting, BiLSTM has an extremely low value of vanishing ratio ($\sim 10^{-3}$) at the 200^{th} training batch (denoted by red vertical line in the plot), when it achieves nearly 100% accuracy on the training data.[7]

Consequently, the BiLSTM model (prematurely) achieves a high training accuracy, solely based on the starting and ending few words, well before the gates can learn to allow the gradients to pass

[5] Further details to aid reproducibility are in the Appendix B.2.

[6] Implementation detail: we choose the left end, as some sequences in a batch might be padded with zeros on the right.

[7] Refer to Appendix C for plots of other pooling techniques.



(a) Gradient Norms (b) BiLSTM (c) MAXATT

Figure 2: (a): The gradient norm ($\|\frac{\partial L}{\partial h_t}\|$) across different word positions. BiLSTM$_{\text{LowF}}$ suffers from extreme vanishing of gradients, with the gradient norm in the middle nearly 10^{-15} times that at the ends. In contrast, pooling methods result in gradients of nearly the same value, irrespective of the word position. (b), (c): The vanishing ratio ($\|\frac{\partial L}{\partial h_{\text{mid}}}\| / \|\frac{\partial L}{\partial h_{\text{end}}}\|$) over training batches for BiLSTM and MAXATT, using 1K, 20K unique training examples from the IMDb dataset. The respective training and validation accuracies are also depicted.

	Vanishing ratio			Validation acc.		
	1K	5K	20K	1K	5K	20K
BiLSTM	5×10^{-3}	0.03	0.06	64.9	82.8	88.4
MEANPOOL	2.5	0.56	1.32	78.4	82.6	88.5
MAXPOOL	0.40	0.42	0.53	78.0	84.7	89.6
ATT	3.87	1.04	1.19	77.1	84.6	90.0
MAXATT	0.69	0.69	0.64	78.1	86.0	90.2

Table 2: Values of vanishing ratio as computed when different models achieve 95% training accuracy, along with the best validation accuracy for that run.

through (and mitigate the vanishing gradients problem). Further reduction in vanishing ratio is unable to improve validation accuracy, due to saturation in training. To examine this more closely, we tabulate the vanishing ratios at the point where the model reaches 95% accuracy on the training data in Table 2. A low value at this point indicates that the gradients are still skewed towards the ends, even as the model begins to overfit on the training data. The vanishing ratio is low for BiLSTM, especially in low-data settings. This results in a 13-14% lower test accuracy in the 1K data setting, compared to other pooling techniques. We conclude that the phenomenon of vanishing gradients results in poorer performance of BiLSTMs. Encouragingly, pooling methods do not exhibit low vanishing ratios, right from the beginning of training, leading to performance gains as demonstrated in the next section.

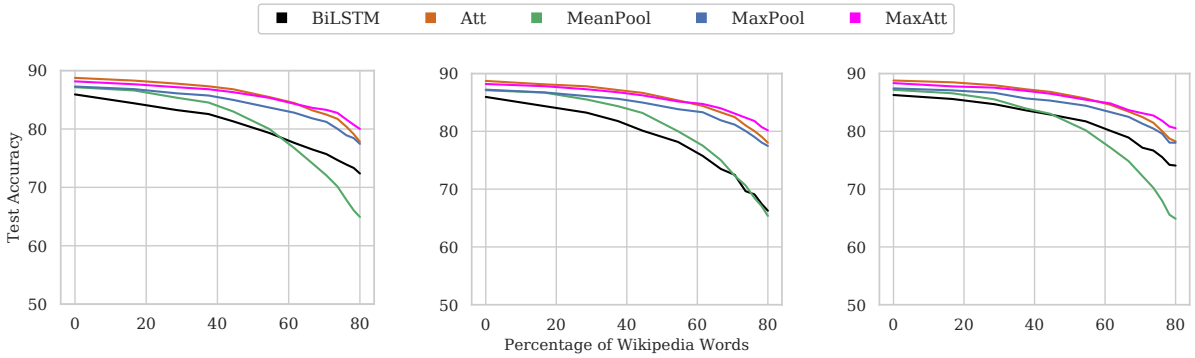
6 Positional Biases

Analyzing the gradient propagation in BiLSTMs suggests that standard recurrent networks are biased towards the end tokens, as the overall contribution of distant hidden states is extremely low in the gradient of the loss. This implies that the weights of various parameters in an LSTM cell (all cells of an LSTM have tied weights) are hardly influenced by the middle words of the sentence. In this light, we aim to evaluate positional biases of recurrent architectures with different pooling techniques.

6.1 Evaluating Natural Positional Biases

Can organically trained recurrent models skip over unimportant words on either ends of the sentence?

Experimental Setup: We append randomly chosen Wikipedia sentences to the input examples of two text classification tasks, based on IMDb and Amazon Reviews, *only at test time*, keeping the training datasets unchanged. Wikipedia sentences are declarative statements of fact, and should not influence the sentiment of movie reviews, and given the diverse nature of the Wikipedia sentences it is unlikely that they would interfere with the few categories (i.e. the labels) of Amazon product reviews. Therefore, it is not unreasonable to expect the models to be robust to such random noise, even though they were not trained for the same. We perform this experiment in three configurations, such that original input is preserved on the (a) left, (b) middle, and (c) right of the modified input. For these configurations, we vary the length of added Wikipedia text in proportion to the length of the original sentence. Figure 4 illustrates the setup when 66% of the total words come from Wikipedia.



(a) Left (b) Mid (c) Right

Figure 3: For models trained on 10K examples, varying amounts of random Wikipedia sentences are appended to the original IMDb reviews *at test time*. Original review is preserved on the (a) left; (b) middle; and (c) right of the modified input. Performance degrades significantly for BiLSTM and MEANPOOL, whereas ATT, MAXPOOL and MAXATT are more resilient.

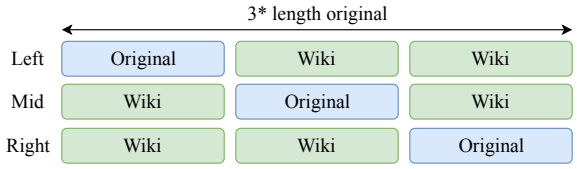


Figure 4: Explaining Wikipedia sentence addition.

Results: The effect of adding random words can be seen in Figure 3. We draw two conclusions: (1) Adding random sentences on both ends is more detrimental to the performance of BiLSTM as compared to the scenario where the input is appended to only one end.[8] This corroborates our previous findings that these models largely account for information at the ends for their predictions. (2) We speculate that paying equal importance to all hidden states prevents MEANPOOL from distilling out important information effectively, making it more susceptible to random noise addition. On the contrary, both max-pooling and attention based architectures like MAXPOOL, ATT and MAXATT are significantly more robust in all the settings. This indicates that max-pooling and attention can help account for salient words and ignore unrelated ones, regardless of their position. Lastly, we provide concurring results on the Amazon dataset, and examine the robustness of different models given lesser training data in Appendix D.

[8]One practical implication of this finding is that adversaries can easily attack middle portions of the input text.

6.2 Training to Skip Unimportant Words

How well can different models be trained to skip unrelated words?

Experimental setup: We create new training datasets by appending random Wikipedia sentences to the original input examples of the datasets described in § 4, such that 66% of the text of each new training example comes from Wikipedia sentences (see Figure 4). We experiment with a varying number of training examples, however, the test set remains the same for fair comparisons.

Results The results are presented in Table 3. First, we note that BiLSTM severely suffers when random sentences are appended at both ends. In fact, the accuracy of BiLSTM in mid settings drops to 50%, 12%, 5%, 50% on IMDb, Yahoo, Amazon, Yelp datasets respectively, which is equal to the majority class baseline. However, the performance drop (while large) is not as drastic when sentences are added to only one end of the text. We speculate that this is because a BiLSTM is composed of a forward and a backward LSTM, and when random sentences are appended to the left, the backward LSTM is able to capture information about the original sentence on the right and vice versa.

Second, while accuracies of all pooling techniques begin to converge given sufficient data, the differences in low training data regime are substantial. Further, the poor performance of BiLSTM re-validates the findings of § 5, where we hypothesize that the model's training saturates before the gradients can learn to reach the middle tokens.[9]

[9]Results on more dataset sizes, and the 'left' setting are in

	IMDb			IMDb (mid) + Wiki			IMDb (right) + Wiki		
	1K	2K	10K	1K	2K	10K	1K	2K	10K
BiLSTM	64.7 ± 2.3	75.0 ± 0.4	86.6 ± 0.8	49.6 ± 0.7	49.9 ± 0.5	50.3 ± 0.3	53.5 ± 2.5	64.7 ± 2.8	85.9 ± 0.5
MEANPOOL	73.0 ± 3.0	81.7 ± 0.7	87.1 ± 0.6	69.8 ± 2.1	76.2 ± 1.0	84.1 ± 0.7	70.0 ± 1.1	76.8 ± 1.0	84.8 ± 0.9
MAXPOOL	69.0 ± 3.9	80.1 ± 0.5	87.8 ± 0.6	64.5 ± 1.8	77.2 ± 2.0	86.0 ± 0.8	65.9 ± 4.6	77.8 ± 0.9	**87.2** ± 0.6
ATT	75.7 ± 2.6	**82.8** ± 0.8	**89.0** ± 0.3	75.0 ± 0.8	79.4 ± 0.8	86.7 ± 1.4	74.7 ± 1.4	80.2 ± 1.8	87.1 ± 1.0
MAXATT	**75.9** ± 2.2	82.5 ± 0.4	88.5 ± 0.5	**75.4** ± 2.4	**80.9** ± 1.8	**86.8** ± 0.5	**77.9** ± 0.9	**81.9** ± 0.5	**87.2** ± 0.5
	Yahoo			Yahoo (mid) + Wiki			Yahoo (right) + Wiki		
	1K	2K	10K	1K	2K	10K	1K	2K	10K
BiLSTM	38.3 ± 4.8	51.4 ± 2.1	63.5 ± 0.6	12.7 ± 1.1	12.7 ± 1.1	11.4 ± 0.8	18.8 ± 2.5	37.3 ± 0.9	60.1 ± 1.5
MEANPOOL	48.2 ± 2.3	56.6 ± 0.5	64.7 ± 0.6	31.9 ± 2.3	43.1 ± 2.0	58.5 ± 0.6	33.9 ± 2.1	43.2 ± 1.0	58.6 ± 0.4
MAXPOOL	50.2 ± 2.1	56.3 ± 1.8	63.9 ± 1.1	33.0 ± 1.0	40.1 ± 1.4	58.4 ± 1.2	33.1 ± 2.5	41.2 ± 0.9	60.9 ± 1.0
ATT	47.3 ± 2.2	54.2 ± 1.1	**65.1** ± 1.5	39.4 ± 0.5	45.1 ± 1.8	61.5 ± 1.7	37.9 ± 1.4	47.6 ± 2.3	62.2 ± 0.9
MAXATT	**51.8** ± 1.1	**57.0** ± 1.1	**65.1** ± 1.1	**39.6** ± 0.9	**48.5** ± 0.6	**62.2** ± 1.6	**40.3** ± 1.5	**50.1** ± 1.6	**63.1** ± 0.7
	Amazon			Amazon (mid) + Wiki			Amazon (right) + Wiki		
	1K	2K	10K	1K	2K	10K	1K	2K	10K
BiLSTM	38.5 ± 4.2	52.7 ± 7.7	76.2 ± 0.7	5.3 ± 0.3	5.4 ± 0.3	5.1 ± 0.4	7.9 ± 0.6	27.9 ± 9.9	70.8 ± 1.5
MEANPOOL	44.8 ± 9.8	55.6 ± 6.4	76.9 ± 0.4	34.4 ± 3.5	52.7 ± 3.5	70.3 ± 1.7	33.3 ± 1.0	48.2 ± 3.4	71.9 ± 0.8
MAXPOOL	49.6 ± 3.9	61.6 ± 2.6	**79.1** ± 0.4	17.0 ± 0.7	34.5 ± 2.0	72.8 ± 0.6	17.0 ± 1.7	36.5 ± 3.0	72.4 ± 0.3
ATT	54.1 ± 5.2	61.2 ± 2.9	77.0 ± 0.3	48.0 ± 1.7	59.1 ± 1.8	**75.3** ± 0.5	48.9 ± 1.5	58.9 ± 1.3	**75.7** ± 0.3
MAXATT	**58.2** ± 3.8	**65.6** ± 0.9	77.3 ± 0.2	**57.7** ± 0.5	**63.0** ± 0.8	74.8 ± 0.5	**57.8** ± 0.8	**63.7** ± 0.8	75.3 ± 0.3
	Yelp			Yelp (mid) + Wiki			Yelp (right) + Wiki		
	1K	2K	10K	1K	2K	10K	1K	2K	10K
BiLSTM	80.7 ± 4.1	84.9 ± 8.0	93.1 ± 0.1	50.2 ± 0.4	51.1 ± 0.9	51.4 ± 0.7	59.4 ± 3.7	79.6 ± 6.2	92.7 ± 0.4
MEANPOOL	**87.1** ± 1.2	**87.9** ± 1.7	93.4 ± 0.3	79.2 ± 1.1	86.7 ± 1.0	92.7 ± 0.2	79.4 ± 0.9	87.1 ± 0.6	92.3 ± 0.4
MAXPOOL	84.4 ± 2.0	86.4 ± 5.1	93.4 ± 0.2	81.1 ± 1.5	85.6 ± 0.6	92.5 ± 0.4	80.6 ± 0.8	86.7 ± 0.9	**93.2** ± 0.2
ATT	82.5 ± 3.7	85.6 ± 6.5	**93.7** ± 0.2	84.4 ± 1.0	89.3 ± 1.0	92.5 ± 0.6	**84.8** ± 0.7	89.1 ± 0.9	92.8 ± 0.4
MAXATT	81.3 ± 5.1	86.0 ± 6.3	**93.7** ± 0.3	**85.1** ± 0.8	**89.4** ± 0.5	**92.9** ± 0.3	84.1 ± 2.5	**89.5** ± 0.7	93.0 ± 0.4

Table 3: Mean test accuracy (± std) (in %) across 5 random seeds. In low-resource settings, MAXATT consistently outpeforms other pooling variants. The performance of different pooling methods converges with increase in data.

Third, when the number of classes is large (as in Yahoo and Amazon datasets), we observe a significant performance difference between ATT and MAXATT. We speculate that as the number of labels increase, a single global query vector (as in ATT) is inadequate to identify important words relevant to each label, whereas a sentence dependent query (as in MAXATT) mitigates this concern.

Evaluation on Short Sentences Finally, we re-evaluate this experiment on (new) datasets with short sentences ($<$ 100 words). Results for the standard and 'mid' settings are presented in Table 4. Unlike long sequences, where BiLSTM model was no better than majority classifier (see Table 3), with shorter sequences, the BiLSTM model performs better. This result supports our hypothesis that the effect of vanishing gradients is prominent in longer sequences.[10] Overall, among all the scenarios discussed in tables 3 and 4, on comparing all pooling methods (and BiLSTM) on the basis of their mean test accuracy, **MAXATT is the best performing model in about 80% cases, ATT in 18% cases.**

	Datasets with Short Sentences			
	Yahoo		Yahoo (mid) + Wiki	
	1K	10K	1K	10K
BiLSTM	20.5 ± 2.9	42.4 ± 0.2	9.9 ± 0.7	24.2 ± 0.9
MEANPOOL	23.1 ± 1.8	43.0 ± 0.3	14.9 ± 2.2	32.8 ± 0.8
MAXPOOL	23.0 ± 2.8	**43.3** ± 0.4	14.1 ± 2.6	33.8 ± 1.2
ATT	24.3 ± 1.1	43.1 ± 0.2	16.9 ± 3.0	37.6 ± 0.5
MAXATT	**25.1** ± 2.2	**43.3** ± 0.3	**18.2** ± 2.4	**37.8** ± 0.8
	Amazon		Amazon (Mid) + Wiki	
	1K	10K	1K	10K
BiLSTM	26.6 ± 4.4	54.0 ± 2.6	5.6 ± 0.4	37.9 ± 0.9
MEANPOOL	29.4 ± 4.0	54.4 ± 2.6	10.8 ± 1.9	46.5 ± 0.5
MAXPOOL	33.5 ± 4.5	55.9 ± 2.0	10.6 ± 1.8	47.0 ± 0.9
ATT	36.4 ± 3.7	55.6 ± 0.6	17.4 ± 3.2	**49.7** ± 0.3
MAXATT	**37.4** ± 3.8	**56.2** ± 0.8	**17.8** ± 4.6	**49.7** ± 0.5

Table 4: Mean test accuracy (± std) (in %) on standard, 'mid' settings across 5 random seeds on Yahoo, Amazon datasets with **short** sentences ($<$ 100 words).

Appendix E.1. Conclusions drawn from the 'right' setting are in line with the observations from the 'left'.

[10] Refer to Appendix E.2 for full evaluation.

6.3 Fine-grained Positional Biases

How does the position of a word affect its contribution to a model's prediction?

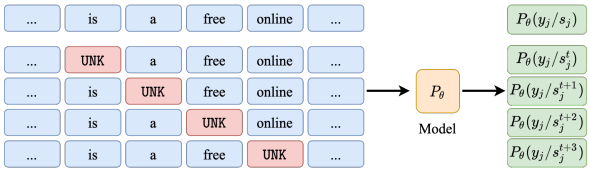


Figure 5: Explaining NWI evaluation.

Experimental Setup: We aim to achieve a fine-grained understanding of model biases w.r.t. each word position, as opposed to evaluating the same at a coarse level (between left, mid and right) as in the previous experiment (§ 6.2). To this end, we define Normalized Word Importance (NWI), a metric to determine the per-position importance of words as attributed by the model. It measures the importance of a particular word (or a set of words) on a model's prediction by calculating the change in the model's confidence in the prediction after replacing it with UNK. (Figure 5). The evaluation is further extended by removing a sequence of k consecutive words to get a smoother metric. The metric is adapted from past efforts to assign word importance, with some differences (Khandelwal et al., 2018; Verwimp et al., 2018; Jain and Wallace, 2019).[11] We provide a complete description of the algorithm to compute NWI in Appendix F, along with further evaluation on IMDb and Amazon datasets.

Results: The results from this experiment are presented in Figure 6 (on the Yahoo dataset). The NWI for architectures with pooling indicate no bias w.r.t. word position, however for BiLSTM there exists a clear bias towards the extreme words on either ends (c.f. Figure 6a). The word importance plots in Figure 6b & 6c demonstrate how pooling is able to *learn* to disambiguate between words that are important for sentence classification significantly better as opposed to BiLSTM. There is a clear peak in the middle in case of 'mid' setting, and on the left in case of 'left' setting for all the pooling architectures. BiLSTM is unable to respond to middle words in Figure 6c. However, they show reasonably higher importance to the left tokens in Figure 6b which is justified by their good performance in the 'left' experimental setting in Table 3. Results for NWI evaluation on all datasets and modified settings (left, mid and right) are available in Appendix F, and are consistent with the representative graphs in Figure 6. We also perform such an analysis on models that are trained on datasets with shorter sentences. Interestingly, the NWI analysis for the Yahoo short dataset in Figure 6d shows that while BiLSTM can better respond to middle words for shorter sentences, it still remains heavily biased towards the ends. We detail these findings in Appendix F.1

[11] Unlike our metric, Khandelwal et al. (2018) remove all words beyond a certain context, and thus capture how important are *all* the removed words, and not one particular word. Jain and Wallace (2019), in their leave-one-out approach, delete a given word rather than replacing it with UNK, thus shifting positions of words by one.

7 Discussion & Conclusion

Through detailed analysis we identify *why* and *when* pooling representations are beneficial in RNNs. While some of the results pertaining to gradient propagation in pooling-based RNNs may be obvious in hindsight, we note that this is the first work to systematically and explicitly analyze the phenomenon.

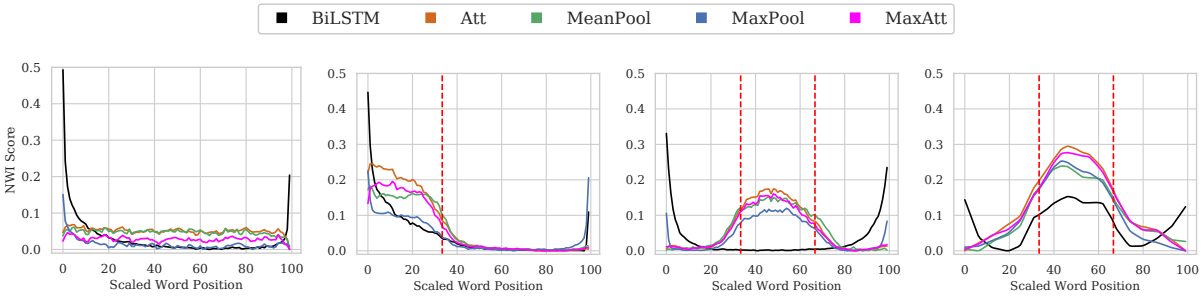
1. We attribute the performance benefits of pooling techniques to their learning ability (*pooling mitigates the problem of vanishing gradients*), and positional invariance (*pooling eliminates positional biases*). Our findings suggest that pooling offers large gains when training examples are few and long, or when salient words lie in the middle of the sequence.

2. In § 5, we observe that gradients in BiLSTM vanish only in initial iterations, but recover slowly during further training. We link the observation with training saturation to provide insights as to why BiLSTMs fail in low-resource setups but pooled architectures don't.

3. We show that BiLSTMs suffer from positional biases even when sentence lengths are as short as 30 words (Figure 6d).

4. We note that pooling makes models significantly more robust to insertions of random words on either end of the input *regardless* of the amount of training data (Figures 3, 8, 9).

5. Lastly, we introduce a novel pooling technique (max-attention) that combines the benefits of max-pooling and attention and achieves superior performance on 80% of our tasks.



(a) Standard (b) Left (c) Mid (d) Mid-Short

Figure 6: Normalized Word Importance w.r.t. word position averaged over examples of length between 400-500 on the Yahoo (25K) dataset in (a,b,c) using $k = 5$; and NWI for examples of length between 50-60 on the Yahoo Short (25K) dataset in (d) with $k. = 3$. Results shown for 'standard', 'left' & 'mid' training settings described in § 6.2. The vertical red line represents a separator between relevant and irrelevant information (by construction).

Most of our insights are derived for sequence classification tasks using RNNs. While our proposed pooling method and analyses are broadly applicable, it remains a part of the future work to evaluate its impact on other tasks and architectures.

Acknowledgemets

We thank Sankalan Pal Chowdhury, Mansi Gupta, Gantavya Bhatt, Atishya Jain, Kundan Krishna and Vishal Sharma for their insightful comments and help with the paper. Mausam is supported by IBM AI Horizons Network for grant, an IBM SUR award, grants by Google, Bloomberg and 1MG, Jai Gupta Chair Fellowship and a Visvesvaraya faculty award by Govt. of India. We thank IIT Delhi HPC facility for computational resources.

References

Martín Arjovsky, Amar Shah, and Yoshua Bengio. 2015. Unitary evolution recurrent neural networks. In *ICML*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate.

Yoshua Bengio, Patrice Y. Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5 2:157–66.

Y-Lan Boureau, Francis Bach, Yann LeCun, and Jean Ponce. 2010a. Learning mid-level features for recognition. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2559–2566. IEEE.

Y-Lan Boureau, Nicolas Le Roux, Francis Bach, Jean Ponce, and Yann LeCun. 2011. Ask the locals: multi-way local pooling for image recognition. In *2011 International Conference on Computer Vision*, pages 2651–2658. IEEE.

Y-Lan Boureau, Jean Ponce, and Yann LeCun. 2010b. A theoretical analysis of feature pooling in visual recognition. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 111–118.

A. P. Sarath Chandar, Chinnadhurai Sankar, Eugene Vorontsov, Samira Ebrahimi Kahou, and Yoshua Bengio. 2019. Towards non-saturating recurrent units for modelling long-term dependencies. In *AAAI*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *EMNLP*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Felix A. Gers, Juergen Schmidhuber, and Fred Cummins. 2000. Learning to forget: continual prediction with lstm. In *Neural Computation*.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256.

S Hochreiter. 1991. Untersuchungen zu dynamischen neuronalen netzen. *Diploma thesis, T.U. Munich*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9:1735–1780.

Alon Jacovi, Oren Sar Shalom, and Yoav Goldberg. 2018. Understanding convolutional neural networks for text classification. *arXiv preprint arXiv:1809.08037*.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation.

Rie Johnson and Tong Zhang. 2016. Supervised and semi-supervised text categorization using lstm for region embeddings. In *ICML*.

Urvashi Khandelwal, He He, Peng Qi, and Daniel Jurafsky. 2018. Sharp nearby, fuzzy far away: How neural language models use context. In *ACL*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *AAAI*.

Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Avinash Madasu and Vijjini Anvesh Rao. 2019. Sequential learning of convolutional features for effective text classification.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2017. Regularizing and optimizing lstm language models. *arXiv preprint arXiv:1708.02182*.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. 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 188–197, Hong Kong, China. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C Lipton. 2020. Learning to deceive with attention-based explanations.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Lyan Verwimp, Hugo Van hamme, Vincent Renkens, and Patrick Wambacq. 2018. State gradients for rnn memory analysis. In *BlackboxNLP@EMNLP*.

Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus. 2013. Regularization of neural networks using dropconnect. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1058–1066, Atlanta, Georgia, USA. PMLR.

Jos van der Westhuizen and Joan Lasenby. 2018. The unreasonable effectiveness of the forget gate.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. *Proceedings of the 2019 conference on Empirical Methods in Natural Language Processing, EMNLP*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016a. 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.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016b. 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, San Diego, California. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 649–657, Cambridge, MA, USA. MIT Press.

Ye Zhang and Byron Wallace. 2015. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*.

Supplementary Material

A Equations for Recurrent Networks

In this section, we provide a mathematical formulation of the equations governing LSTMs and basic RNNs.

A.1 Basic RNN

Recurrent Neural Networks use a series of input sequence x_t and pass it sequentially over a network of hidden states where each each hidden state leads to the next. Mathematically, this is given by:

$$h_t = \sigma(Ux_t + Wh_{t-1} + b)$$
$$y_t = \text{softmax}(Vh_t + c)$$

where x_t refers to the input sequence at time step t, and W, U, V are weights for the RNN cell, and σ is a non-linearity of choice.

A.2 LSTM

The forward propagation of information in a basic LSTM are governed by the following equations:

$$i_t = \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi})$$
$$f_t = \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf})$$
$$g_t = \tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg})$$
$$o_t = \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho})$$
$$c_t = f_t * c_{t-1} + i_t * g_t$$
$$h_t = o_t * \tanh(c_t)$$

where at time t, h_t is the hidden state, c_t is the cell state, x_t is the input, and i_t, f_t, g_t, o_t are the input, forget, cell, and output gates, respectively. σ is the sigmoid function, and * is the Hadamard product.

A.3 MeanPool

For a mean-pooled LSTM, while the forward propagation remains the same as BiLSTM, the output embedding is given by:

$$s^i_{emb} = \frac{\sum_{t\in(1,n)} h^i_t}{n}$$

where h^i_t represents the i^{th} dimension of the hidden state at time step = t, and s_{emb} represents the final output embedding returned by the recurrent structure. This implies that during backpropagation we find a direct influence of the t^{th} hidden state as:

$$\frac{\partial s^i_{emb}}{\partial h^i_t} = \frac{\sum_{k\in(1,n)} \frac{\partial h^i_k}{\partial h^i_t}}{n}$$

B Datasets and Experimental Settings

B.1 Dataset Extraction

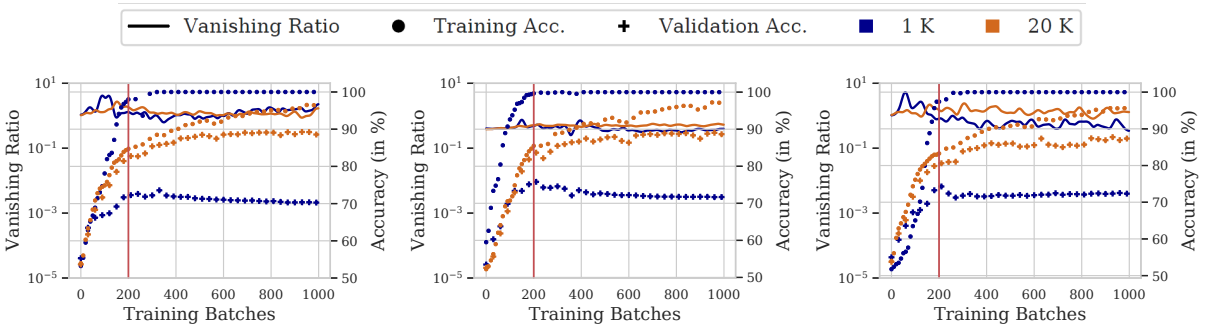
Amazon Reviews The Amazon Reviews Dataset (Ni et al., 2019) includes reviews (ratings, text, helpfulness votes) and product metadata (descriptions, category etc.) pertaining to products on the Amazon website. We extract the product category and review text corresponding to 2500 reviews from to each of the following 20 classes:

- Automotive
- Books
- Clothing Shoes and Jewelry
- Electronics
- Movies and TV
- Arts Crafts and Sewing
- Toys and Games
- Pet Supplies
- Sports and Outdoors
- Grocery and Gourmet Food
- CDs and Vinyl
- Tools and Home Improvement
- Software
- Office Products
- Patio Lawn and Garden
- Home and Kitchen
- Industrial and Scientific
- Luxury Beauty
- Musical Instruments
- Kindle Store

In the standard setting, we ensure that all reviews have lengths between 100 and 500 words.

IMDb The IMDb Movie Reviews Dataset (Maas et al., 2011) is a popular binary sentiment classification task. We take a subset of 20000 reviews that have length greater than 100 words for the purposes of experimentation in this paper.

Yahoo Yahoo! Answers (Zhang et al., 2015) has over 1,400,000 question and answer pairs spread across 10 classes. We do not use information such as question, title, date and location for the purpose of classification. As in the case of Amazon reviews, in the standard setting, we ensure that all answers have lengths between 100 and 1000 words, while in the short sentence setting, the maximum answer length in the filtered dataset is 100 words.



(a) ATT (b) MAXPOOL (c) MEANPOOL

Figure 7: The vanishing ratio ($\|\frac{\partial L}{\partial h_{\text{end}}}\| / \|\frac{\partial L}{\partial h_{\text{mid}}}\|$) over training steps for ATT, MAXPOOL, MEANPOOL using 1K, 20K training examples from the IMDb dataset. The respective training and validation accuracies are also depicted.

Yelp Reviews : Yelp Reviews (Zhang et al., 2015) is a sentiment analysis task with 5 matching classes. For the purposes of experimentation, we create a subset which is filtered to contain sentences in the range 100 to 1000 tokens. Further, all reviews with a score of 4 or 5 are maked positive, while those with a score of 1 or 2 are marked negative for the binary classification task.

B.2 Reproducibility

Computing Infrastructure For all the experiments described in the paper, we use a Tesla K40 GPUs supporting a maximum of 10GB of GPU memory. All experiments can be performed on a single GPU. The brief experimentation done on transformer models was done using Tesla V100s that support 32 GB of GPU memory.

Run Time The average run-time for each epoch varies linearly with the amount of training data and average sentence length. For the mode with 25K training data in standard setting (sentences with greater than 100 words, and no wikipedia words) the average training time for 1 epoch is under 2 minutes. Further, across all pooling techniques, the run time varies only marginally.

Number of Parameters The number of parameters in the model varies with the vocabulary size. We cap the maximum vocabulary size to 25,000 words. However, in the 1K training data setting, the actual vocabulary size is lesser (depending on the training data). The majority of the parameters of the model are accounted for in the model's embedding matrix = (vocabulary size)×(embedding size). The number of parameters for the main LSTM model are around 70,000, with the ATT model having a few more parameters than other methods due to a learnable query vector.

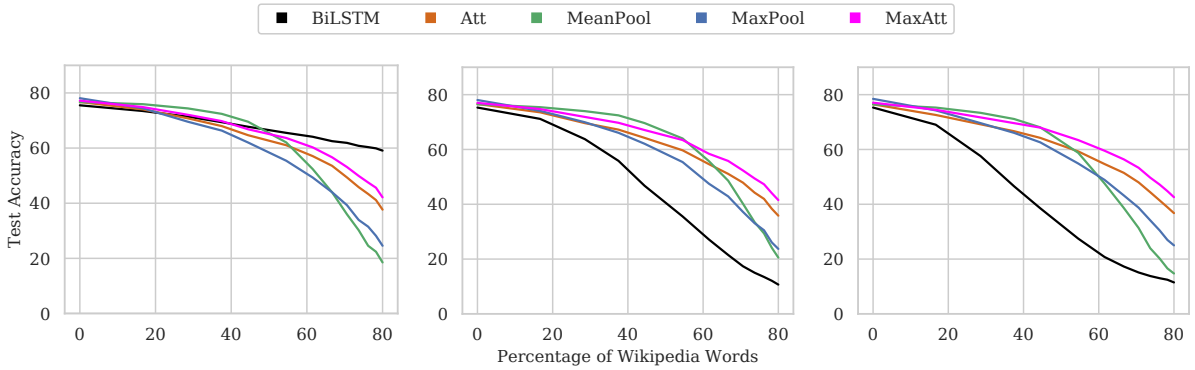
Validation Scores We provide validation results in Table 2 for the standard setting. However, in interest of brevity, we only detail the test scores in all subsequent tables. Note that we always select the model based on the best validation accuracy during the training process (among all the epochs).

Evaluation Metric The evaluation metric used is the model's accuracy on the test set and is reported as an average over 5 different seeds. All the classes are nearly balanced in the datasets chosen, hence standard accuracy metric serves as an accurate indicator.

Hyperparameters search An explicit hyperparameter search is not performed for each model in each training setting over all seeds, since the purpose of the paper is not to beat the state of art, but rather to analyze the effect of pooling in recurrent architectures. We do note that, in the manual search performed on the learning rates of $\{1 \times 10^{-3}, 2 \times 10^{-3}, 5 \times 10^{-3}\}$ on the IMDb and Yahoo datasets, we find that for all the pooling and non-pooling methods discussed, models trained on learning rate equal to 2×10^{-3} showed the best validation accuracy. Thus, we use that for all the following results. However, we do perform a hyperparameter search for the best regularization parameters as described in Appendix E.3. We keep the embedding dimension and hidden dimension fixed for all experiments.

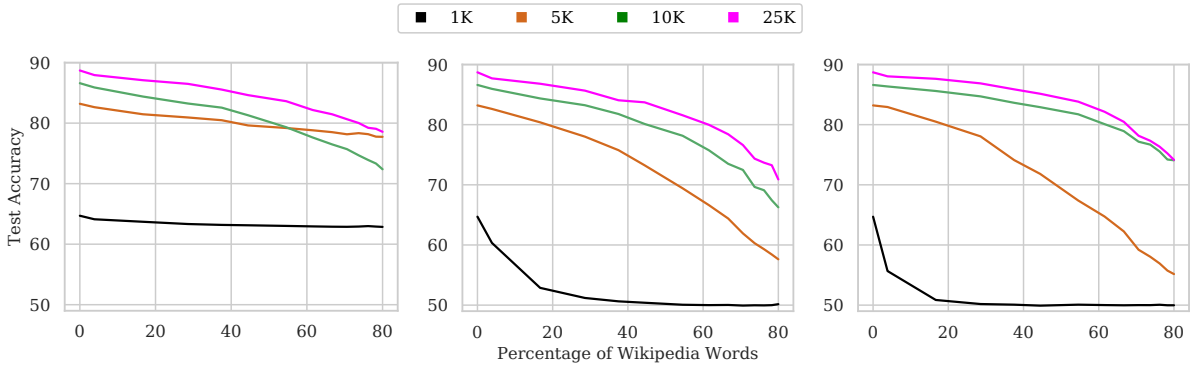
C Gradient Propagation

The plots of the change in vanishing ratios for ATT, MAXPOOL and MEANPOOL are shown in Figure 7.



(a) Left (b) Mid (c) Right

Figure 8: Amazon Dataset (10K setting): Random Wikipedia sentences are appended to the original input paragraphs. Original input is preserved on the (a) left, (b) middle, and (c) right of the new input. Test accuracies are reported by varying the percentage of total Wikipedia words in the new input.



(a) Left (b) Mid (c) Right

Figure 9: IMDb Dataset (BiLSTM): Random Wikipedia sentences are appended to the original input paragraphs for the standard BiLSTM models trained on 1K, 5K, 10K and 20K examples. Original input is preserved on the (a) left, (b) middle, and (c) right of the new input. Test accuracies are reported by varying the percentage of total Wikipedia words in the new input. BiLSTM is unrepsonsive to any appended tokens as long as the 'left' text is preserved in the 1K and 5K setting. But this bias dilutes with more training samples. Given sufficient data (more than 10K unique examples) the effect of appending random words on both ends is more detrimental than that on appending at only one end.

This completes the representative analysis for BiLSTM and MAXATT shown in Figure 2. It can be seen that for all the different pooling types discussed in this paper, the vanishing ratios are small right from the beginning of training. This motivates future research to further formally analyze and discover other learning advantages (apart from vanishing ratios) that distinguish the performance of one pooling technique from the other.

D Evaluating Natural Positional Biases

In line with our results in § 6.1, we further evaluate models trained on the Amazon dataset in the same settings to re-validate our results. The effect of appending random Wikipedia sentences to input examples on models trained on the Amazon dataset can be found in Figure 8. We use the model trained on 10K examples to perform this experiment. The graphs show similar findings as in Figure 3, and further supports the hypothesis that BiLSTM gives a strong emphasis on extreme words when trained on standard datasets, which is why its performance significantly deteriorates when random Wikipedia sentences are appended on both ends.

Effect of Amount of Training Data: Figure 3 suggests that BiLSTM is equally responsive to the effect of appending random words to the left or right. However, in case of the Amazon Reviews dataset (Figure 8), we notice that the BiLSTM is more resilient when the text to the left is preserved.

	Datasets with Long Sentences									
	Yahoo Dataset					Amazon Dataset				
	1K	2K	5K	10K	25K	1K	2K	5K	10K	25K
BiLSTM	38.3 ± 4.8	51.4 ± 2.1	57.4 ± 0.6	63.5 ± 0.6	67.5 ± 0.8	38.5 ± 4.2	52.7 ± 7.7	70.0 ± 0.9	76.2 ± 0.7	81.8 ± 0.3
MEANPOOL	48.2 ± 2.3	56.6 ± 0.5	60.8 ± 0.5	64.7 ± 0.6	68.7 ± 0.6	44.8 ± 9.8	55.6 ± 6.4	71.2 ± 0.9	76.9 ± 0.4	82.0 ± 0.3
MAXPOOL	50.2 ± 2.1	56.3 ± 1.8	61.3 ± 0.9	63.9 ± 1.1	67.0 ± 1.1	49.6 ± 3.9	61.6 ± 2.6	73.9 ± 0.2	79.1 ± 0.4	84.2 ± 0.2
ATT	47.3 ± 2.2	54.2 ± 1.1	61.0 ± 0.5	65.1 ± 1.5	68.2 ± 0.7	54.1 ± 5.2	61.2 ± 2.9	72.0 ± 0.2	77.0 ± 0.3	82.6 ± 0.1
MAXATT	51.8 ± 1.1	57.0 ± 1.1	63.2 ± 0.4	65.1 ± 1.1	68.4 ± 0.6	58.2 ± 3.8	65.6 ± 0.9	72.8 ± 0.5	77.3 ± 0.2	82.4 ± 0.2
	Yahoo (left) + Wiki					Amazon (left) + Wiki				
	1K	2K	5K	10K	25K	1K	2K	5K	10K	25K
BiLSTM	41.4 ± 2.9	51.0 ± 0.5	56.2 ± 1.2	60.9 ± 0.7	64.6 ± 2.0	44.9 ± 0.7	57.0 ± 0.8	68.3 ± 0.9	73.5 ± 0.4	79.6 ± 0.2
MEANPOOL	31.9 ± 1.5	43.3 ± 1.7	51.4 ± 0.9	58.8 ± 0.7	65.1 ± 0.3	31.0 ± 2.1	48.1 ± 1.4	65.0 ± 1.4	70.8 ± 1.2	79.1 ± 0.8
MAXPOOL	33.6 ± 0.9	42.3 ± 1.4	52.7 ± 2.0	60.7 ± 0.9	66.0 ± 1.0	19.2 ± 1.9	42.5 ± 3.5	68.5 ± 2.6	76.8 ± 0.6	82.1 ± 0.5
ATT	37.3 ± 0.5	47.2 ± 2.2	57.6 ± 1.6	62.5 ± 1.0	67.6 ± 0.3	47.6 ± 2.0	59.3 ± 1.1	70.8 ± 0.9	75.6 ± 0.3	81.3 ± 0.3
MAXATT	40.0 ± 0.6	48.7 ± 0.5	59.6 ± 1.4	63.0 ± 1.4	67.2 ± 0.9	56.1 ± 1.3	63.8 ± 1.3	70.3 ± 0.3	75.6 ± 0.2	80.7 ± 0.5
	Yahoo (mid) + Wiki					Amazon (mid) + Wiki				
	1K	2K	5K	10K	25K	1K	2K	5K	10K	25K
BiLSTM	12.7 ± 1.1	12.7 ± 1.1	12.0 ± 0.9	11.4 ± 0.8	13.2 ± 2.2	5.3 ± 0.3	5.4 ± 0.3	5.0 ± 0.1	5.1 ± 0.4	7.8 ± 5.2
MEANPOOL	31.9 ± 2.3	43.1 ± 2.0	50.1 ± 1.6	58.5 ± 0.6	64.9 ± 0.7	34.4 ± 3.5	52.7 ± 3.5	63.4 ± 2.0	70.3 ± 1.7	79.0 ± 0.6
MAXPOOL	33.0 ± 1.0	40.1 ± 1.4	51.0 ± 1.2	58.4 ± 1.2	65.5 ± 0.7	17.0 ± 0.7	34.5 ± 2.0	58.8 ± 0.4	72.8 ± 0.6	80.4 ± 0.3
ATT	39.4 ± 0.5	45.1 ± 1.8	57.0 ± 2.0	61.5 ± 1.7	66.5 ± 0.6	48.0 ± 1.7	59.1 ± 1.8	69.5 ± 0.6	75.3 ± 0.5	81.1 ± 0.2
MAXATT	39.6 ± 0.9	48.5 ± 0.6	58.7 ± 1.5	62.2 ± 1.6	66.5 ± 0.7	57.7 ± 0.5	63.0 ± 0.8	69.8 ± 0.6	74.8 ± 0.5	80.3 ± 0.4
	Yahoo (right) + Wiki					Amazon (right) + Wiki				
	1K	2K	5K	10K	25K	1K	2K	5K	10K	25K
BiLSTM	18.8 ± 2.5	37.3 ± 0.9	52.9 ± 2.1	60.1 ± 1.5	65.4 ± 0.6	7.9 ± 0.6	27.9 ± 9.9	45.8 ± 16.2	70.8 ± 1.5	78.7 ± 0.8
MEANPOOL	33.9 ± 2.1	43.2 ± 1.0	50.6 ± 0.8	58.6 ± 0.4	64.6 ± 0.5	33.3 ± 1.0	48.2 ± 3.4	64.1 ± 0.7	71.9 ± 0.8	78.8 ± 0.2
MAXPOOL	33.1 ± 2.5	41.2 ± 0.9	53.0 ± 3.6	60.9 ± 1.0	66.0 ± 0.7	17.0 ± 1.7	36.5 ± 3.0	64.3 ± 1.5	72.4 ± 0.3	80.2 ± 0.9
ATT	37.9 ± 1.4	47.6 ± 2.3	58.1 ± 1.4	62.2 ± 0.9	67.0 ± 0.3	48.9 ± 1.5	58.9 ± 1.3	69.7 ± 0.6	75.7 ± 0.3	81.1 ± 0.3
MAXATT	40.3 ± 1.5	50.1 ± 1.6	59.3 ± 1.2	63.1 ± 0.7	66.8 ± 0.3	57.8 ± 0.8	63.7 ± 0.8	71.1 ± 0.6	75.3 ± 0.3	80.7 ± 0.5

Table 5: Mean test accuracy (± std) (in %) on different manipulated settings across 5 random seeds on the Yahoo, Amazon datasets with long sentences (greater than 100 words).

This indicates a learning bias, where the BiLSTM pays greater emphasis to outputs of one chain of the bidirectional LSTM. It is interesting to note that on reducing the training data, this bias increases significantly in the case of IMDb dataset as well.

We hypothesize that such a phenomenon may have resulted due to an artifact of the training process itself, that is, the model is able to find 'easily identifiable' important sentiment at the beginning of the reviews during training (speculatively due to the added effects of padding to the right). Therefore, given less training data, BiLSTMs prematurely learn to use features from only one of the two LSTM chains and (in this case) the left $\rightarrow$ right chain of the dominates the final prediction. We confirm from Figure 9 that with a decrease in training data (such as in the 1K IMDb data setting), the bias towards one end substantially increases, that is, BiLSTM is extremely insensitive to random sentence addition, as long as the left end is preserved.

Practical Implications We observe that MEANPOOL and BiLSTM can be susceptible to *changes in test-time data distribution*. This questions the use of such models in real word settings. We speculate that paying equal importance to all hidden states handicaps MEANPOOL from being able to distil out important information effectively, while the preceding discussion on the effect of size of training data highlights the possible cause of this occurrence in BiLSTM. We observe that other pooling methods like MAXATT are able to circumvent this issue as they are only mildly affected by the added Wikipedia sentences.

E Training to Skip Unimportant Words

We demonstrate in § 6.2 that the ability of BiLSTM, and its different pooling variants, to learn to skip unrelated words can be greatly diminished in challenging datasets especially given less amount of input data. In this section, we aim to (a) provide a complete evaluation on all positions of data modi-

Datasets with Long Sentences										
	IMDb Dataset					Yelp Dataset				
	1K	2K	5K	10K	20K	1K	2K	5K	10K	25K
BiLSTM	64.7 ± 2.3	75.0 ± 0.4	83.2 ± 0.4	86.6 ± 0.8	88.7 ± 0.6	80.7 ± 4.1	84.9 ± 8.0	92.2 ± 0.3	93.1 ± 0.1	94.1 ± 0.3
MEANPOOL	73.0 ± 3.0	81.7 ± 0.7	85.4 ± 0.1	87.1 ± 0.6	88.6 ± 0.3	87.1 ± 1.2	87.9 ± 1.7	92.2 ± 0.4	93.4 ± 0.3	94.4 ± 0.1
MAXPOOL	69.0 ± 3.9	80.1 ± 0.5	85.7 ± 0.2	87.8 ± 0.6	89.9 ± 0.3	84.4 ± 2.0	86.4 ± 5.1	92.2 ± 0.3	93.4 ± 0.2	94.7 ± 0.2
ATT	75.7 ± 2.6	82.8 ± 0.8	86.9 ± 0.7	89.0 ± 0.3	90.3 ± 0.2	82.5 ± 3.7	85.6 ± 6.5	92.6 ± 0.4	93.7 ± 0.2	94.9 ± 0.1
MAXATT	75.9 ± 2.2	82.5 ± 0.4	86.1 ± 0.8	88.5 ± 0.5	89.9 ± 0.2	81.3 ± 5.1	86.0 ± 6.3	92.6 ± 0.2	93.7 ± 0.3	94.8 ± 0.1
	IMDb (left) + Wiki					Yelp (left) + Wiki				
	1K	2K	5K	10K	20K	1K	2K	5K	10K	25K
BiLSTM	67.6 ± 1.1	74.7 ± 1.2	80.6 ± 0.3	84.5 ± 0.4	87.2 ± 0.4	81.7 ± 0.5	87.5 ± 0.5	90.7 ± 0.5	92.0 ± 0.3	93.8 ± 0.2
MEANPOOL	69.7 ± 3.4	76.6 ± 0.9	81.7 ± 0.7	83.6 ± 1.0	86.5 ± 0.8	78.1 ± 1.3	87.0 ± 1.1	90.9 ± 0.3	92.5 ± 0.1	93.8 ± 0.2
MAXPOOL	68.8 ± 1.2	76.8 ± 1.7	82.2 ± 0.8	86.9 ± 0.9	88.4 ± 0.5	80.2 ± 1.5	87.5 ± 1.0	91.4 ± 0.2	93.0 ± 0.4	94.3 ± 0.1
ATT	76.5 ± 1.5	79.6 ± 1.1	82.6 ± 0.6	86.9 ± 0.8	88.9 ± 0.5	84.7 ± 1.6	89.5 ± 0.7	92.0 ± 0.2	92.9 ± 0.4	94.4 ± 0.2
MAXATT	75.8 ± 1.5	80.6 ± 1.0	84.1 ± 1.5	87.1 ± 0.6	89.1 ± 0.2	84.7 ± 1.3	89.7 ± 0.6	92.1 ± 0.1	93.1 ± 0.4	94.2 ± 0.4
	IMDb (mid) + Wiki					Yelp (mid) + Wiki				
	1K	2K	5K	10K	20K	1K	2K	5K	10K	25K
BiLSTM	49.6 ± 0.7	49.9 ± 0.5	50.2 ± 0.3	50.3 ± 0.3	50.1 ± 0.3	50.2 ± 0.4	51.1 ± 0.9	51.2 ± 0.8	51.4 ± 0.7	51.5 ± 0.5
MEANPOOL	69.8 ± 2.1	76.2 ± 1.0	82.2 ± 0.7	84.1 ± 0.7	86.5 ± 0.8	79.2 ± 1.1	86.7 ± 1.0	90.7 ± 0.3	92.7 ± 0.2	94.0 ± 0.1
MAXPOOL	64.5 ± 1.8	77.2 ± 2.0	82.9 ± 1.2	86.0 ± 0.8	88.4 ± 0.6	81.1 ± 1.5	85.6 ± 0.6	90.7 ± 0.4	92.5 ± 0.4	94.1 ± 0.2
ATT	75.0 ± 0.8	79.4 ± 0.8	83.4 ± 1.0	86.7 ± 1.4	88.8 ± 0.2	84.4 ± 1.0	89.3 ± 1.0	91.8 ± 0.6	92.5 ± 0.6	94.4 ± 0.2
MAXATT	75.4 ± 2.4	80.9 ± 1.8	84.7 ± 1.3	86.8 ± 0.5	88.7 ± 0.4	85.1 ± 0.8	89.4 ± 0.5	91.7 ± 0.7	92.9 ± 0.3	94.3 ± 0.2
	IMDb (right) + Wiki					Yelp (right) + Wiki				
	1K	2K	5K	10K	20K	1K	2K	5K	10K	25K
BiLSTM	53.5 ± 2.5	64.7 ± 2.8	79.7 ± 4.3	85.9 ± 0.5	88.5 ± 0.2	59.4 ± 3.7	79.6 ± 6.2	91.7 ± 0.3	92.7 ± 0.4	93.7 ± 0.4
MEANPOOL	70.0 ± 1.1	76.8 ± 1.0	81.8 ± 0.5	84.8 ± 0.9	87.1 ± 0.3	79.4 ± 0.9	87.1 ± 0.6	90.9 ± 0.7	92.3 ± 0.4	93.8 ± 0.3
MAXPOOL	65.9 ± 4.6	77.8 ± 0.9	84.9 ± 0.8	87.2 ± 0.6	89.3 ± 0.3	80.6 ± 0.8	86.7 ± 0.9	91.9 ± 0.5	93.2 ± 0.2	94.5 ± 0.3
ATT	74.7 ± 1.4	80.2 ± 1.8	84.7 ± 1.1	87.1 ± 1.0	89.4 ± 0.3	84.8 ± 0.7	89.1 ± 0.9	92.0 ± 0.4	92.8 ± 0.4	94.3 ± 0.1
MAXATT	77.9 ± 0.9	81.9 ± 0.5	85.2 ± 0.8	87.2 ± 0.5	89.4 ± 0.3	84.1 ± 2.5	89.5 ± 0.7	91.7 ± 0.9	93.0 ± 0.4	94.3 ± 0.1

Table 6: Mean test accuracy (± std) (in %) on different manipulated settings across 5 random seeds on the IMDb, Yelp Reviews datasets with long sentences (less than 100 words).

fication and dataset size settings (including those which were skipped in the main paper for brevity); (b) evaluate the same experiment in a setting where input examples are shorter in length.

E.1 Full Evaluation

For completeness, we perform the evaluation in § 6.2 on each of {1K, 2K, 5K, 10K, 25K} dataset size settings, and also report the results when Wikipedia words are appended on the right, preserving the original input to the left. We report results for the Yahoo and Amazon datasets in Table 5 and the IMDb and Yelp Reviews datasets in Table 6. It can be noted that the advantages of MAXATT over other pooling and non-pooling techniques significantly increase in the three Wikipedia settings in each of the tables. This suggests that MAXATT performs better in more challenging scenarios where the important signals are hidden in the input data. Further, the performance advantages of MAXATT are more when amount of training data is less.

E.2 Short Sentences

Dataset	Classes	Avg. Length	Max Length	Train Size	Test Size
Yahoo! Answers	10	30.1	95	25K	25K
Amazon Reviews	20	29.1	100	25K	12.5K

Table 7: Corpus statistics for classification tasks (short datasets).

For shorter sequences, we reuse two of our text classification tasks: (1) **Yahoo! Answers**; and (2) **Amazon** Reviews. Similar to the setting with long sentences in the main paper, we use only the text and labels, ignoring any auxiliary information (like title or location). We select subsets of the datasets with sequences having a length (number of space separated words) less than 100. A summary of statistics with respect to sentence length and corpus size is given in Table 7.

The results for the performance of the trained models can be found in Table 8. In the 'Mid' set-

	Datasets with Short Sentences									
	Yahoo Dataset					Amazon Dataset				
	1K	2K	5K	10K	25K	1K	2K	5K	10K	25K
BiLSTM	20.5 ± 2.9	25.8 ± 3.7	33.1 ± 2.4	42.4 ± 0.2	46.0 ± 0.4	26.6 ± 4.4	37.7 ± 3.4	48.6 ± 2.2	54.0 ± 2.6	61.7 ± 0.2
MEANPOOL	23.1 ± 1.8	28.4 ± 1.5	35.3 ± 1.8	43.0 ± 0.3	46.5 ± 0.4	29.4 ± 4.0	38.2 ± 3.4	49.0 ± 1.8	54.4 ± 2.6	62.0 ± 0.2
MAXPOOL	23.0 ± 2.8	31.2 ± 1.4	37.3 ± 1.9	43.3 ± 0.4	46.8 ± 0.8	33.5 ± 4.5	41.4 ± 3.3	50.8 ± 1.7	55.9 ± 2.0	62.8 ± 0.1
ATT	24.3 ± 1.1	30.7 ± 2.5	36.3 ± 2.0	43.1 ± 0.2	46.4 ± 0.6	36.4 ± 3.7	43.3 ± 1.7	50.9 ± 0.6	55.6 ± 0.6	61.9 ± 0.2
MAXATT	25.1 ± 2.2	30.8 ± 2.6	37.9 ± 1.1	43.3 ± 0.3	46.8 ± 0.7	37.4 ± 3.8	44.6 ± 1.2	51.6 ± 0.8	56.2 ± 0.8	62.4 ± 0.4
	Yahoo (left) + Wiki					Amazon (left) + Wiki				
	1K	2K	5K	10K	25K	1K	2K	5K	10K	25K
BiLSTM	19.6 ± 1.7	28.5 ± 0.8	34.5 ± 0.3	38.2 ± 0.7	43.0 ± 0.4	20.0 ± 1.8	30.7 ± 1.9	43.4 ± 0.4	49.9 ± 0.4	56.1 ± 0.3
MEANPOOL	17.0 ± 2.9	20.3 ± 0.3	29.1 ± 1.1	34.8 ± 1.1	42.0 ± 0.3	10.4 ± 2.1	18.0 ± 2.4	34.3 ± 2.4	46.2 ± 1.1	55.0 ± 0.5
MAXPOOL	15.7 ± 0.8	24.0 ± 1.2	33.5 ± 0.4	37.5 ± 1.0	43.7 ± 0.1	12.4 ± 1.9	26.0 ± 0.6	44.5 ± 1.0	51.4 ± 0.3	57.5 ± 0.2
ATT	19.8 ± 3.1	26.0 ± 0.5	35.5 ± 0.9	40.1 ± 0.5	43.8 ± 0.2	21.3 ± 4.3	37.1 ± 0.7	46.1 ± 0.6	51.3 ± 0.8	57.2 ± 0.2
MAXATT	19.7 ± 3.5	27.0 ± 0.9	36.2 ± 1.3	40.0 ± 0.6	43.7 ± 0.3	22.1 ± 5.9	36.7 ± 1.3	46.7 ± 0.1	52.2 ± 0.1	57.5 ± 0.2
	Yahoo (mid) + Wiki					Amazon (mid) + Wiki				
	1K	2K	5K	10K	25K	1K	2K	5K	10K	25K
BiLSTM	9.9 ± 0.7	12.3 ± 0.8	17.4 ± 1.1	24.2 ± 0.9	36.3 ± 0.5	5.6 ± 0.4	6.9 ± 0.5	20.3 ± 1.0	37.9 ± 0.9	51.5 ± 1.0
MEANPOOL	14.9 ± 2.2	22.1 ± 1.3	28.3 ± 0.4	32.8 ± 0.8	39.2 ± 0.4	10.8 ± 1.9	20.8 ± 1.3	39.0 ± 0.6	46.5 ± 0.5	54.8 ± 0.1
MAXPOOL	14.1 ± 2.6	22.6 ± 0.3	28.6 ± 0.5	33.8 ± 1.2	40.1 ± 0.5	10.6 ± 1.8	21.3 ± 1.7	37.1 ± 1.4	47.0 ± 0.9	55.3 ± 0.4
ATT	16.9 ± 3.0	24.8 ± 1.1	31.4 ± 0.9	37.6 ± 0.5	42.1 ± 0.4	17.4 ± 3.2	33.2 ± 1.0	43.9 ± 0.5	49.7 ± 0.3	55.4 ± 0.1
MAXATT	18.2 ± 2.4	25.7 ± 0.5	32.6 ± 0.6	37.8 ± 0.8	42.1 ± 0.4	17.8 ± 4.6	35.0 ± 1.2	44.7 ± 0.3	49.7 ± 0.5	55.8 ± 0.4
	Yahoo (right) + Wiki					Amazon (right) + Wiki				
	1K	2K	5K	10K	25K	1K	2K	5K	10K	25K
BiLSTM	12.3 ± 0.5	23.8 ± 1.2	33.4 ± 0.6	38.2 ± 0.2	43.8 ± 0.3	7.4 ± 0.8	15.3 ± 3.2	40.8 ± 0.5	50.4 ± 0.7	58.4 ± 0.4
MEANPOOL	15.7 ± 1.9	22.7 ± 0.4	27.7 ± 0.9	34.2 ± 0.6	41.3 ± 0.1	14.8 ± 2.0	20.4 ± 3.3	40.1 ± 1.2	48.6 ± 0.5	56.9 ± 0.3
MAXPOOL	14.7 ± 0.6	22.5 ± 1.5	33.6 ± 0.5	38.5 ± 0.4	43.4 ± 0.5	11.1 ± 2.3	24.0 ± 1.9	45.6 ± 0.5	52.0 ± 0.4	58.4 ± 0.3
ATT	19.7 ± 0.2	27.4 ± 1.5	35.9 ± 0.2	40.0 ± 0.4	43.8 ± 0.7	22.4 ± 5.7	36.6 ± 1.3	46.7 ± 0.4	52.5 ± 0.4	59.1 ± 0.3
MAXATT	20.3 ± 1.3	28.1 ± 0.9	35.4 ± 0.8	40.3 ± 0.4	43.8 ± 0.4	20.8 ± 6.8	37.3 ± 0.9	47.8 ± 0.4	53.1 ± 0.3	59.0 ± 0.2

Table 8: Mean test accuracy (± std) (in %) on different manipulated settings across 5 random seeds on the Yahoo, Amazon datasets with short sentences (less than 100 words).

ting, we observe that BiLSTM performs significantly better on shorter sequences as opposed to the long sequences. For instance, in case of Amazon Dataset (Mid), under the 25K data setting, the classification accuracy increases from 7.8% in Table 5 to 51.5% in Table 8, which is a significant improvement from only doing as well as majority guessing in the former. We note that most of the learning issues of BiLSTM in long sentence setting are largely absent when sentence lengths are short, with BiLSTM also emerging as the best-performing model in a few cases. This corroborates the effect of gradients vanishing with longer time steps.

E.3 On using regularization

For the experiments in the work, we do not regularize trained LSTMs. This has two analytical advantages (1) we can examine the benefits of pooling without having to account for the the effect of regularization; and (2) training to 100% accuracy acts as an indicator of training the models *adequately*. However, for validation, we also performed our experiments on the IMDb dataset with 2 different types of regularization schemes, following best practices used in previous works (Merity et al., 2017). We use DropConnect (Wan et al., 2013) [12] and Weight Decay [13] for regularization of all the models. We observe that the effect of regularization consistently improves the final accuracies by 1-2% across the board. However, even after sustained training (up to 50 epochs), BiLSTM still suffers from the learning issues outlined in the paper. The goal of this paper is not to study the effect of various regularization schemes, but to merely understand the effect pooling in improving the performance of BiLSTM.

F Fine-grained Positional Biases

We detail the method for calculating the Normalized Word Importance (NWI) score in Algorithm 1.

[12] grid search over mask rate: {0.1,0.3,0.5}

[13] grid search over decay value: $\{10^{-3}, 10^{-4}, 10^{-6}, 10^{-8}\}$

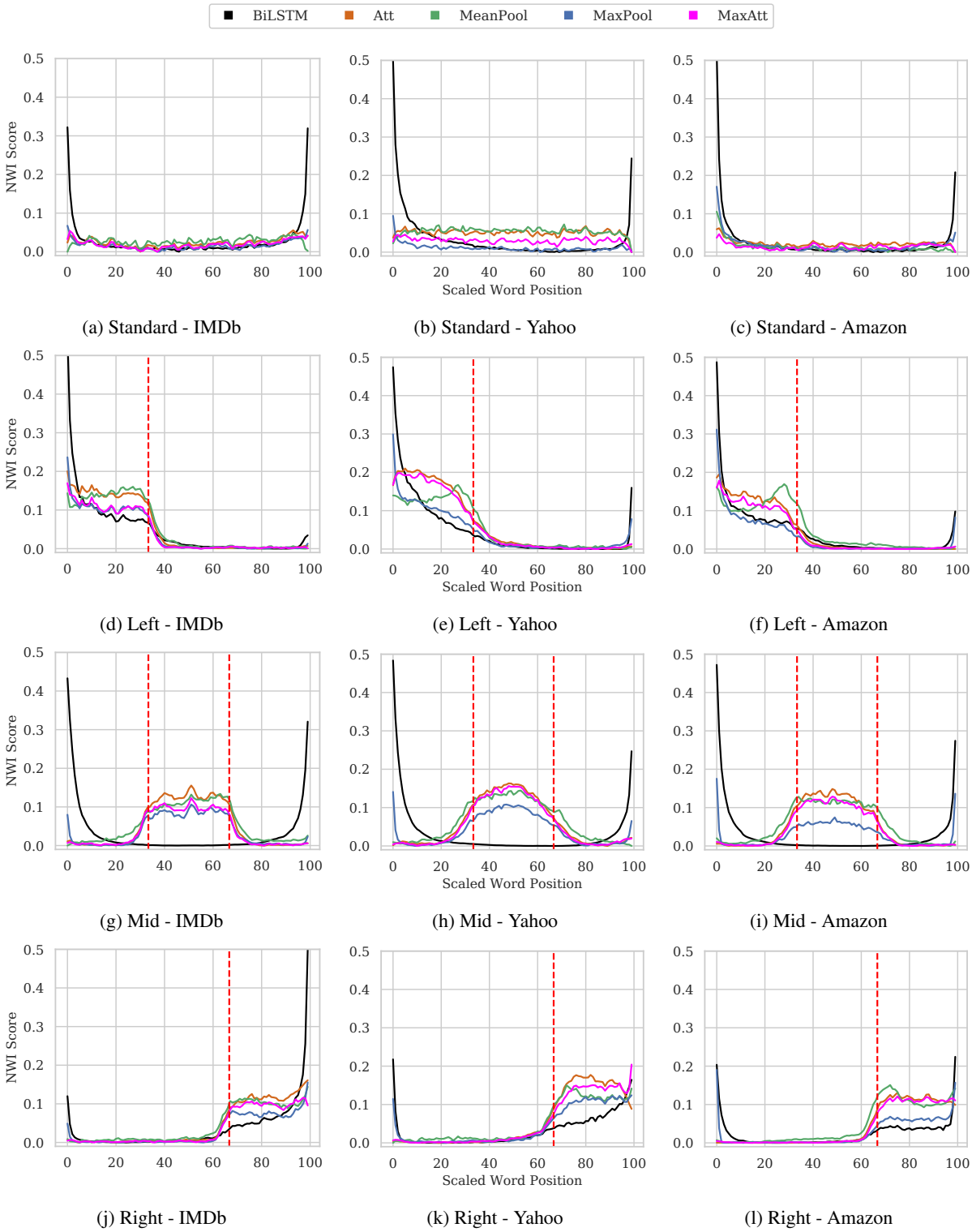
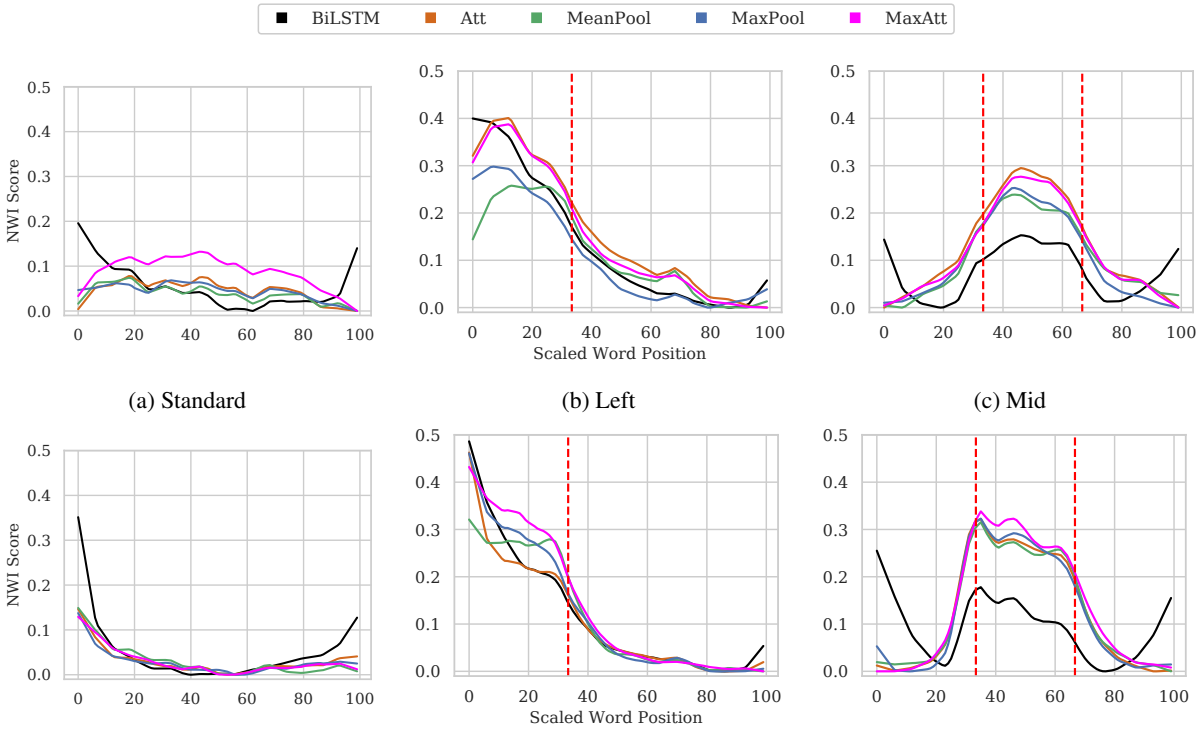


Figure 10: Normalized Word Importance w.r.t. word position for k = 5; averaged over sentences of length between 400-500 on the IMDb, Yahoo, Amazon (10K) Datasets. Results shown for the 'standard', 'left', 'mid' and 'right' training settings described in § 6.2. The vertical red line represents an approximate separator between relevant and irrelevant information (by construction). For instance, The word positions to the 'left' of the vertical line in graphs in the second row of the Figure contain data from true input examples, while those to the right contain Wikipedia sentences.



(a) Standard (b) Left (c) Mid

(d) Standard (e) Left (f) Mid

Figure 11: Normalized Word Importance w.r.t. word position for k = 3; averaged over sentences of length between 50-60 on the Yahoo, Amazon (10K) Datasets. Results shown for the 'standard', 'left' and 'mid' training settings described in Appendix E.2. The vertical red line represents an approximate separator between relevant and irrelevant information (by construction). For instance, The word positions to the 'left' of the vertical line in (b), (e) contain data from true input examples, while those to the right contain Wikipedia sentences.

Algorithm 1 *NWI* evaluation

Input: softmax classifier P_θ, test set D
Parameters: k
for $s_j = \{x_j^1, \ldots, x_j^n\}, y_j$ in D **do**
  $p_j = \log\{P_\theta(y_j|s_j)\}$
  for $t = 0 \ldots \frac{n}{k}$ **do**
    $s_j^t = \{x_j^1, \ldots, x_j^{k.t}, \underbrace{\texttt{UNK}, \ldots, \texttt{UNK}}_{\text{k words}}, \ldots, x_j^n\}$
    $p_j^t = \log\{P_\theta(y_j|s_j^t)\}$
    $\delta_j^t = |p_j^t - p_j|$
  end for
  $\text{nwi}_j = \frac{\delta_j}{\max_{t\in(1,\frac{n}{k})}\delta_j^t}$
  $\text{nwi}_j = \text{nwi}_j - \min_{t\in(1,\frac{n}{k})}\delta_j^t$
  $\text{nwi}_j = \texttt{LinInterp}(\text{nwi}_j, \frac{n}{k}, 100)$
end for
return $\frac{1}{|D|}\sum_{j=1}^{|D|}\text{nwi}_j$
*$\texttt{LinInterp}(x, n, l)$ linearly interpolates input distribution x of n discrete steps to l steps.

The parameter k can be adjusted according to the average sentence length. For a sentence of length 100, setting an extremely low value of k (say 1) may have very little impact of the model's prediction $\log\{P_\theta(y_j|s_j^t)\}$ for all positions t. On the other hand, setting an extremely high value of k (say 20) may provide only few data points, and also change the model prediction drastically at all values of t.

Complete graphs for the positional importance (as perceived by the model) of words are detailed in this section. The trends observed for the remaining datasets are similar to the representative graphs shown in the main paper. We show graphs for the IMDb, Yahoo and Amazon datasets in Figure 10.

Practical Implications Our findings suggest that adversaries can easily replace the middle portion of texts with racist or abusive sentences, and still stay undetected by BiLSTM based detection systems. This is because BiLSTM attributes little or no importance to words in the middle of the input. Pooling based models are able to circumvent this issue by being able to attribute importance to words irrespective of their position.

F.1 NWI for Short sentences

We repeat our experiments of NWI evaluation on the datasets with short sentences ($<$100 words) as described in Appendix E.2. It is interesting to observe the graphs on the Yahoo and Amazon short datasets in Figure 11, where due to the short sentence length, even BiLSTM is able to show the desired importance characteristic in case of mid setting. This supports the fact that the test time accuracies in the mid setting are no longer as bad as a majority class predictor. Interestingly, in case of short sentences in the mid setting (Figures 11c,11f), we observe three peaks in the NWI graph. The one in the middle is expected given the data distribution. However, the two peaks in NWI at the extreme ends help establish that while BiLSTM is able to propagate gradients to the middle given the short sentences, it is still not able to forego the extreme bias towards the end tokens.