# Finding the Needle in a Haystack: Unsupervised Rationale Extraction from Long Document Classifiers

**Anonymous ACL submission**

## Abstract

Long-sequence models are designed to better represent longer texts and improve performance on document-level tasks. Advancing from individual sentences to using much longer context, rationale extraction from these models is becoming increasingly important, in order to analyse model behaviour and provide finer-grained informative predictions. This paper investigates methods of unsupervised rationale extraction for long-sequence models in the context of document classification. We find that previously proposed methods for sentence classification do not perform well when applied on long documents, due to very limited tokens being updated during training. We alleviate this issue by introducing a Ranked Soft Attention architecture that ensures more tokens receive appropriate weak supervision. We also investigate a Compositional Soft Attention architecture that applies RoBERTa sentence-wise to extract plausible rationales at the token-level. The proposed methods significantly outperform Longformer-driven baselines on sentiment classification datasets, while also exhibiting significantly lower runtimes.

## 1 Introduction

Transformer-based architectures (Vaswani et al., 2017) have become ubiquitous in natural language processing research. A key attribute to their success is the multi-head self-attention mechanism (Michel et al., 2019). However, its computational and memory requirements grow quadratically with input sequence length. Therefore, models such as BERT (Devlin et al., 2019) commonly limit the maximum sequence length to 512 tokens. Longer documents are truncated (Devlin et al., 2019) or staggered position embeddings are used (Jain et al., 2020). This limitation motivated long-text transfomers, such as Big Bird (Zaheer et al., 2020) and Longformer (Beltagy et al., 2020), which reduce the complexity of self-attention through the use of sparse attention and improve the performance of transformers on long documents.

The task of rationale extraction focuses on selecting a subset of input as a justification for the model's output. The extracted tokens can then be used for verification of document-level predictions (Dzindolet et al., 2003) or as input to another model (Jain et al., 2020). In particular, our work focuses on extracting token-level rationales from long document models. The architectures are optimised using only document labels, with no token-level annotations used during training. This setting is more practical for longer texts, where token-level annotations are often missing due to prohibitive costs of manual labeling.

Existing work has investigated rationale extraction for regular transformer-based classifiers (Pruthi et al., 2020; Jain et al., 2020; Bujel et al., 2021; Fomicheva et al., 2022). However, to the best of our knowledge, there is no work that investigates unsupervised rationale extraction for long-text transformers. As we show, methods designed for standard transformers do not necessarily perform as well on longer documents. We highlight that this work focuses on extracting *plausible* rationales (agreeable to human annotators; DeYoung et al. (2020)), as opposed to *faithful* explanations (true to the system's computation; Rudin (2018)).

We investigate various methods to adapt long-text transformers to zero-shot rationale extraction. We find that the sparse self-attention present in Longformer struggles to locate tokens that can serve as a plausible rationale for document labels. For longer datasets, the quality of the rationale extracted from the self-attention layers is not better than a random baseline. Following a qualitative investigation, we propose Ranked Soft Attention and Compositional Soft Attention architectures, which significantly improve over other unsupervised approaches for rationale extraction on longer sentiment detection documents.

## 2 Soft Attention

Rei and Søgaard (2018) introduced a soft attention architecture for biLSTM zero-shot sequence labelers, which Bujel et al. (2021) adapted to transformers by introducing Weighted Soft Attention. We apply Weighted Soft Attention to contextual embeddings $t_i \in \mathbb{R}^h$ for all tokens in the document:

$$e_i = \tanh(W_e t_i + b_e) \quad (1)$$

$$\tilde{a}_i = \sigma(\tilde{e}_i) \qquad \tilde{e}_i = W_{\tilde{e}} e_i + b_{\tilde{e}} \quad (2)$$

where $h \in \mathbb{N}$ is the dimension of the contextual token embeddings, $e_i \in \mathbb{R}^{h'}$ is a hidden layer, $\tilde{e}_i \in \mathbb{R}$ is a single scalar value, $\sigma$ is the sigmoid activation function and $\tilde{a}_i \in [0, 1]$ is the token attention score. The scores are converted to normalized attention weights $a_i$ to build document-level representations $c$:

$$a_i = \frac{\tilde{a}_i^{\beta}}{\sum_{j=1}^{N} \tilde{a}_j^{\beta}} \qquad c = \sum_{i=1}^{N} a_i t_i \quad (3)$$

$$d = \tanh(W_d c + b_d) \qquad y = \sigma(W_y d + b_y) \quad (4)$$

where $N$ is the number of tokens in a sentence, $c \in \mathbb{R}^h$ is the final document representation, $d \in \mathbb{R}^s$ is the hidden document representation and $y \in \mathbb{R}$ is the document-level prediction. $\beta \in \mathbb{R}$ is a weight controlling the sharpness of the attention scores. For each document $j \in \mathbb{N}$, we obtain document-level predictions $y^{(j)} \in [0, 1]$ and token-level scores $0 \leq \tilde{a}_i \leq 1$. As the token-level scores are calculated using a logistic activation function, we use a classification threshold of $0.50$. The supervision of Weighted Soft Attention is described in Appendix A.

### 2.1 Ranked Soft Attention

Empirically, we find that most of the token scores obtained using Weighted Soft Attention are close to 1 for long documents. We suspect this is caused by an insufficient number of token scores receiving supervision signal at each epoch – only minimum and maximum token scores are optimized by the soft attention method. While working well for individual sentences, such an approach does not scale to longer documents. We find that, on average, only 5% of tokens receive supervision signal during training.

We instead propose Ranked Soft Attention, where $k\%$ of tokens with the highest scores are supervised with the document label, while the remaining $100 - k\%$ tokens are supervised with a 0 label. This is achieved through the following loss function $L_{ranked}$:

$$L_{top} = \sum_{j}^{M} \sum_{\tilde{a}_i \in \tilde{A}_j} I_{top}(\tilde{a}_i)(\tilde{a}_i - \tilde{y}^{(j)})^2 \quad (5)$$

$$L_{rest} = \sum_{j}^{M} \sum_{\tilde{a}_i \in \tilde{A}_j} (1 - I_{top}(\tilde{a}_i))(\tilde{a}_i - 0)^2 \quad (6)$$

$$L_{ranked} = \frac{1}{k} L_{top} + \frac{1}{100 - k} L_{bottom} \quad (7)$$

where $M$ is the number of documents, $\tilde{A}_j$ is the set of tokens in document $j$, $I_{top}(\tilde{a}_i)$ is an indicator function that returns 1 if $\tilde{a}_i$ is in the top $k\%$ of token scores and 0 otherwise, and $k$ is a hyperparameter that can be tuned based on the percentage of annotations in the dataset (Jain et al., 2020). $L_{top}$ steers the top token scores towards 1 for the positive documents only. This ensures the model only provides rationale for the positive class texts. $L_{rest}$ encourages the model to ignore low-scoring tokens by assigning them scores close to 0. This term enforces sparsity of the token scores, ensuring that only a subset of tokens are weighted highly. Unlike Weighted Soft Attention, this setup ensures all tokens receive supervision signal. The total loss is then $L = L_{doc} + \gamma_{ranked} L_{ranked}$, where $L_{doc}$ is the document-level loss.

### 2.2 Compositional Soft Attention

Given RoBERTa's strong performance as a rationale extractor for individual sentences (Bujel et al., 2021; Pruthi et al., 2020), we investigate the feasibility of using RoBERTa together with Ranked Soft Attention to extract token-level rationales from longer documents. However, due to RoBERTa's standard 512 sequence length limit, a direct application to longer texts is not feasible. Instead, we propose to model intra-sentence token dependencies by applying RoBERTa to each sentence individually. To extract rationales for the whole document, we use a Ranked Soft Attention layer that composes the individual contextual token embeddings across different sentences. This is in contrast to Hierarchical Transformers (Pappagari et al., 2019), which focused on document-level representation without obtaining token-level rationale.

Our proposed Compositional Soft Attention architecture uses a standard length transformer to

| | IMDb-Pos | | | | | IMDb-Neg | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Doc $F_1$ | $F_1$ | $F_{0.5}$ | MAP | Time | Doc $F_1$ | $F_1$ | $F_{0.5}$ | MAP | Time |
| Random Uniform | - | 5.46 | 4.27 | 8.46 | - | - | 6.02 | 4.80 | 9.90 | - |
| Longformer Weighted Soft Attention | 93.41 | 6.89 | 4.43 | 7.81 | 603 | 93.52 | 8.34 | 5.40 | 10.59 | 599 |
| Longformer Self-Attention Top-K | **94.42** | 5.90 | 5.19 | 7.85 | 606 | **94.60** | 5.71 | 5.02 | 9.03 | 603 |
| Longformer Ranked Soft Attention | 92.63 | 14.13 | 11.92 | 11.97 | 618 | 93.92 | 19.62 | 16.98 | 16.44 | 608 |
| Compositional Soft Attention | 91.85 | **25.46** | **20.12** | **26.82** | **436** | 90.82 | **27.27** | **22.70** | **29.78** | **433** |

Table 1: Results on the Sentiment Detection IMDb dataset. Doc $F_1$ represents document-level classification performance, while $F_1$, $F_{0.5}$ and *MAP* are token-level metrics. *Time* represents the average seconds per epoch each model took to train. We find our Compositional Soft Attention to perform best at the token-level rationale extraction.

build contextual token embeddings $s'_k \in \mathbb{R}^{N_k \times h}$ separately for each sentence $s_k$:

$$s'_k = \text{Transformer}(s_k) \tag{8}$$

$$t = \text{Concat}[s'_0, ..., s'_{m_j}] \tag{9}$$

where $m_j$ is the number of sentences in document $j$ with $0 < k \leq m_j$, $h$ is the output dimension of the transformer, $N_k \in \mathbb{R}$ is the number of tokens in sentence $s_k$ and $t \in \mathbb{R}^{N \times h}$ contains the concatenated representations of all tokens in the document.

We provide this document representation $t$ as input to our Ranked Soft Attention layer (Section 2.1), which composes tokens across all sentences to obtain a document-level representation and the prediction $y^{(j)}$. We additionally obtain token-level attention scores $\tilde{a}_i$, which we use to extract rationales for document classification tasks. An architecture visualization and pseudocode are given in Appendix B.

## 3 Datasets

We investigate the performance of our models on three different datasets. Each of the datasets contains a document-level label, together with human annotations on the token-level that are used for evaluation only.

We evaluate our models on Grammatical Error Detection (GED) datasets, which contain texts written by non-native learners of English. They are annotated with token-level grammatical errors, which serve as rationale for document-level proficiency scores. Specifically, the **BEA 2019**[1] shared task (Bryant et al., 2019) released a set of essays from Write & Improve, an online automated assessment and feedback platform (Yannakoudakis et al., 2018). The essays were submitted in response to various prompts, and document-level labels indicate the CEFR proficiency level (A/B/C)[2]. We remove intermediate (B) essays, treat the beginner (A) class as the positive documents and the advanced (C) class as the negative label. This is motivated by the beginner class containing a higher proportion of grammatically incorrect tokens, which we treat as rationale for the low proficiency level. Since there is no publicly available test set, we held out the development set for evaluation and randomly sample 10% of the training data for development.

The First Certificate in English[3] **(FCE)** dataset (Yannakoudakis et al., 2011) contains essays written for an intermediate-level language proficiency exam. Each student wrote 2 essays, which have been given an overall exam script score. We treat these concatenated essays as one document. We split the dataset into beginner (score $< 27$; equivalent to a fail) and advanced learners (score $> 30$), and use the train/dev/test split released by Rei and Yannakoudakis (2016). We note that both GED datasets contain a relatively small number of documents exceeding 512 tokens, which is the standard RoBERTa maximum sequence length (Table 3). We use them due to lack of longer datasets with document-level and token-level annotations.

We also use the Sentiment Detection movie reviews IMDb[4] dataset (Zaidan et al., 2007), which contains positive and negative movie reviews. We focus on a subset of this dataset that has been annotated with rationales for the reviews by human annotators. We split the dataset into **IMDb-Pos** and **IMDb-Neg**, where the former contains evidence for only positive reviews, while the latter contains evidence for negative ones. We use the train/dev/test split published by Pruthi et al. (2020).

[1] https://www.cl.cam.ac.uk/research/nl/bea2019st/

[2] https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions

[3] https://ilexir.co.uk/datasets/index.html

[4] https://www.tensorflow.org/datasets/catalog/movie_rationales

| | FCE | | | | | BEA 2019 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Doc $F_1$ | $F_1$ | $F_{0.5}$ | MAP | Time | Doc $F_1$ | $F_1$ | $F_{0.5}$ | MAP | Time |
| Random Uniform | - | 12.13 | 13.70 | 15.91 | - | - | 11.36 | 12.00 | 16.68 | - |
| Longformer Weighted Soft Attention | **89.34** | **24.81** | 17.64 | 16.17 | 242 | 93.79 | **21.22** | 14.59 | 16.38 | 357 |
| Longformer Self-Attention Top-K | 89.29 | 14.23 | 15.57 | 14.67 | 243 | **94.39** | 14.75 | 14.99 | 13.76 | 360 |
| Longformer Ranked Soft Attention | 89.10 | 22.23 | **23.40** | 18.82 | 238 | 93.49 | 19.11 | 19.65 | 18.55 | 351 |
| Compositional Soft Attention | 81.32 | 21.08 | 23.19 | **19.67** | **139** | 89.07 | 20.44 | **20.22** | **19.44** | **141** |

Table 2: Results for the Grammatical Error Detection Datasets. Doc $F_1$ represents document-level classification performance, while $F_1$, $F_{0.5}$ and *MAP* are token-level metrics. *Time* represents the average seconds per epoch each model took to train. We note that Weighted Soft Attention achieves a high $F_1$ due to assigning 1 to most tokens, thus resulting in high recall. Our Compositional Soft Attention is on-par with Longformer-based models.

## 4 Results

We report $F_1$ and $F_{0.5}$ on the token level. As the rationale proportion in our datasets is low, the $F_{0.5}$ metric allows us to better distinguish between models that display high recall but low precision and more well-balanced ones. We further evaluate performance using Mean Average Precision (MAP) for ranking positive tokens. MAP removes the threshold dependency and indicates which models return the best token ranking. We perform significance testing using a two-tailed paired t-test ($a = 0.05$).

As a baseline, we use the Longformer Self-Attention Top-K (Jain et al., 2020), where we select the top $k\%$ of the global attention scores from the <CLS> token in the Longformer's last layer. These high ranking tokens are rationale if the document-level prediction is the positive class. Details of our experimental setup are presented in Appendix D. Tables 1 and 2 present results for the GED and Sentiment Detection datasets respectively. We provide example token-level predictions in Appendix F.

We find that both Longformer Self-Attention and Longformer Weighted Soft Attention perform poorly on the task of rationale extraction. On the longer Sentiment Detection dataset, the performance of both methods is on-par with a random baseline. For GED, we note that the higher token-level $F_1$ score of Weighted Soft Attention is due to the model assigning scores of 1 to most tokens, as evident by the substantially lower $F_{0.5}$ score. We suspect this homogeneity of token scores to be caused by only 2 tokens per document receiving supervision signal at each epoch. This encourages the scores to stay close to the initial values, meaning the model is unable to learn to provide plausible rationales for its predictions. Increasing the supervision signal through Ranked Soft Attention significantly improves the token-level $F_{0.5}$ ($5.06\% - 11.58\%$ absolute increase).

We suspect the poor performance of Longformer Self-Attention is partly due to the use of normalized global attention from the <CLS> token. As the token attention scores across the whole document have to sum up to 1, very few tokens are assigned high scores. This is evidenced by the significantly lower token-level recall compared to other methods. While some improvement could be obtained by fine-tuning the classification threshold, self-attention methods do not return a good token ranking, as indicated by the low MAP scores, which are below the random baseline.

Compositional Soft Attention significantly outperforms all models on the longer IMDb datasets and achieves results on-par with Longformer Ranked Soft Attention on the GED datasets. The compositional nature of this architecture allows RoBERTa to learn to provide meaningful token-level scores for each sentence individually. This is in contrast to other Longformer-based methods, which focus on modeling global dependencies and struggle to provide good token-level predictions. We also notice that Compositional Soft Attention, overall, exhibits substantially lower runtimes ($30\% - 60\%$) than the Longformer methods.

## 5 Conclusion

We investigated unsupervised rationale extraction for long document classifiers. Our experiments showed that standard Transformer-based soft attention methods do not perform well on longer texts. We proposed Ranked Soft Attention that works well with Longformer by increasing the supervision signal available to individual tokens. We further introduced a novel Compositional Soft Attention architecture that extends RoBERTa to represent long documents. We found Compositional Soft Attention to significantly outperform Longformer-based systems on rationale extraction for longer documents, while being $30\% - 60\%$ faster to fine-tune.

## Limitations

Our work aims to fill a gap in the literature for effective rationale extraction from long text classifiers. We improve the scalability of such methods by sequentially applying RoBERTa to each sentence instead of relying on slower long text transformers. However, it is important to underline that this method still does not permit application to texts of arbitrary length, as the memory of the GPU is the main limitation. We hope to address this in future work.

We also note the limited evaluation for truly long documents. This is due to the small number of long text datasets with token-level annotations available. We encourage the development of such new datasets in the future.

We believe that current approaches of framing rationale extraction as a sequence labeling problem do not scale well to longer documents. It is becoming difficult to quantitatively evaluate such models without taking into accounts spans of predictions and the annotations. We encourage future work to investigate alternative methods of evaluating plausibility of token-level rationale extractors for long texts.

## Ethics Statement

The Grammatical Error Detection datasets were obtained from responses to various prompts either online or during an English exam. We note that both datasets were annotated by experts with knowledge of the CEFR language proficiency scale. As minors constitute a large group of English learners, it is likely that a part of responses come from them. However, we note that the authors of the original datasets obtained all necessary consents and anonymized the essays. The dataset contains no harmful language. FCE dataset is released under a non-commercial research license and for educational purposes. BEA 2019 is similarly available under a non-commercial license.

On the other hand, the IMDb dataset annotation was crowd-sourced from Amazon Mechanical Turk workers. While the authors put measures in-place to ensure consistency of the annotations, we note that they were likely performed by non-experts and thus might be less comprehensive than the Grammatical Error Detection datasets. The dataset contains no personally identifiable information, but may contain offensive language. The dataset is released under a personal and non-commercial license.

We note that our models do not provide explanations of their decisions. They focus on extracting rationale that is plausible to a human annotator, instead of a faithful explanation. These two terms can be easily confused by a non-expert reader and can lead to incorrect application of the architecture.

## References


Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.

Kamil Bujel, Helen Yannakoudakis, and Marek Rei. 2021. Zero-shot sequence labeling for transformer-based sentence classifiers. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 195–205, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. 2003. The role of trust in automation reliance. *International Journal of Human-Computer Studies*, 58(6):697–718.

Marina Fomicheva, Lucia Specia, and Nikolaos Aletras. 2022. Translation error detection as rationale extraction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4148–4159.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. 2020. Learning to faithfully rationalize by construction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4459–4473, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis,

Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? *Advances in Neural Information Processing Systems*, 32.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2020. On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations*.

Raghavendra Pappagari, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. Hierarchical transformers for long document classification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 838–844. IEEE.

Danish Pruthi, Bhuwan Dhingra, Graham Neubig, and Zachary C Lipton. 2020. Weakly-and semi-supervised evidence extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3965–3970.

Marek Rei and Anders Søgaard. 2018. Zero-shot sequence labeling: Transferring knowledge from sentences to tokens. 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 293–302.

Marek Rei and Helen Yannakoudakis. 2016. Compositional sequence labeling models for error detection in learner writing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1181–1191.

Cynthia Rudin. 2018. Please stop explaining black box models for high stakes decisions. *stat*, 1050:26.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Helen Yannakoudakis, Øistein E Andersen, Ardeshir Geranpayeh, Ted Briscoe, and Diane Nicholls. 2018. Developing an automated writing placement system for ESL learners. *Applied Measurement in Education*, 31(3):251–267.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 180–189.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big Bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the main conference*, pages 260–267.

## A Supervision of Weighted Soft Attention

We further recall that the Weighted Soft Attention architecture uses the following loss functions:

$$L_{doc} = \sum_j (y^{(j)} - \tilde{y}^{(j)})^2 \quad (10)$$

$$L_1 = \sum_j (min(\tilde{a}_i) - 0)^2 \quad (11)$$

$$L_2 = \sum_j (max(\tilde{a}_i) - \tilde{y}^{(j)})^2 \quad (12)$$

$$L = L_{doc} + \gamma(L_1 + L_2) \quad (13)$$

where $L_{doc}$ optimizes the document-level performance, $L_1$ ensures the minimum attention score is close to 0 and $L_2$ optimizes the maximum attention score to be close to the document label $\tilde{y}^{(j)}$.

## B Compositional Soft Attention

We present the algorithm for Compositional Soft Attention in Algorithm 1 and the overview of the architecture in Figure 1.

**Algorithm 1** Compositional Soft Attention

**for** sentence $s_{j,i}$ in document$_j$ **do**
  $s'_{j,i} \leftarrow$ Transformer$(s_{j,i})$
**end for**
$t_j \leftarrow [s'_{j,1}, ..., s'_{j,m_j}]$
$y^{(j)},\ \tilde{a} =$ SoftAttention$(t_j)$

## C Datasets

We present the summary statistics for the datasets used in our experiments in Table 3.

## D Experimental Setup

We use a pre-trained RoBERTa-base (Liu et al., 2019) and Longformer-base (Beltagy et al., 2020) models, available through the HuggingFace library (Wolf et al., 2020). All experiments are performed on Nvidia Tesla P100. Following Mosbach et al. (2020), we train for 20 epochs, with the best performing checkpoint chosen. Each experiment is repeated 3 times and the average results are reported. We set $k$ based on the percentage of evidence present for the positive class. We use $\gamma_{ranked} = 1.0$ and $\gamma = 1.0$ where appropriate. The Longformer-based models have approximately $1.49 \times 10^8$ trainable parameters, while the Compositional Soft Attention has $1.25 \times 10^8$ trainable parameters.

## E Full Results

We present full results of the experiments in Tables 4, 5, 6 and 7.

## F Example Predictions

Furthermore, we provide more sample predictions made by different models in Figures 2 and 3.

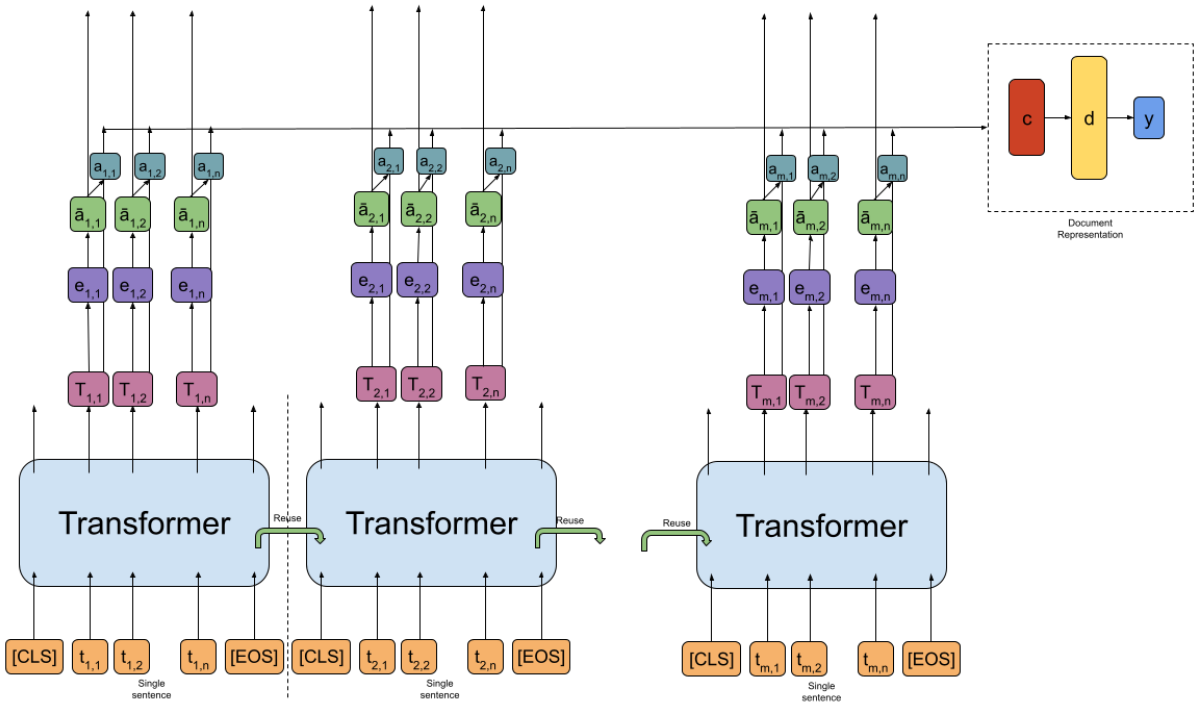


Figure 1: Compositional Soft Attention architecture for rationale extraction and document classification. A standard size transformer is applied to each sentence individually and the contextual token embeddings are then combined to build a document-level representation.

| | FCE | BEA 2019 | IMDb-Pos | IMDb-Neg |
|---|---|---|---|---|
| Number of train samples | 722 | 1120 | 1200 | 1200 |
| Number of dev samples | 51 | 280 | 299 | 299 |
| Number of test samples | 66 | 200 | 300 | 300 |
| Average text length (words) | 441 | 213 | 686 | 686 |
| Maximum text length (words) | 725 | 655 | 1935 | 1935 |
| % of texts $>$ 512 words | 16% | 2% | 73% | 73% |
| % positive samples | 49% | 46% | 50% | 50% |
| % negative samples | 51% | 54% | 50% | 50% |
| % evidence | 13% | 9% | 8% | 8% |

Table 3: Statistics for the datasets used. All measured on the development datasets. We note the low proportion of long texts in FCE and BEA 2019.

| | Doc $F_1$ | $F_1$ | $F_{0.5}$ | $P$ | $R$ | MAP |
|---|---|---|---|---|---|---|
| Random Uniform | - | $11.36 \pm 0.32$ | 12.00 | 12.47 | 10.42 | 16.68 |
| Longformer Weighted Soft Attention | 93.79 | **21.22** $\pm 0.17$ | 14.59 | 12.08 | **88.38** | 16.38 |
| Longformer Soft Attention Top-K | **94.39** | $14.75 \pm 3.21$ | 14.99 | 15.16 | 14.40 | 13.76 |
| Longformer Ranked Soft Attention | 93.49 | $19.11 \pm 2.92$ | 19.65 | **20.12** | 18.53 | 18.55 |
| Compositional Soft Attention | 89.07 | $20.44 \pm 1.73$ | **20.22** | 20.10 | 20.95 | **19.44** |

Table 4: Full results for BEA 2019. We note that while Weighted Soft Attention performs best on the token-level, it is largely due to the model assigning scores of 1 to most tokens, as indicated by the high recall. Using $F_{0.5}$ as an evaluation metric highlights this issue. Our proposed Compositional Soft Attention performs best on the token-level in terms of both $F_{0.5}$ and MAP.

| | Doc $F_1$ | $F_1$ | $F_{0.5}$ | $P$ | $R$ | MAP |
|---|---|---|---|---|---|---|
| Random Uniform | - | $12.13 \pm 0.36$ | 13.70 | 14.99 | 10.18 | 15.91 |
| Longformer Weighted Soft Attention | **89.34** | **24.81** $\pm 1.42$ | 17.64 | 14.81 | **83.35** | 16.17 |
| Longformer Soft Attention Top-K | 89.29 | $14.23 \pm 2.25$ | 15.57 | 16.61 | 12.46 | 14.67 |
| Longformer Ranked Soft Attention | 89.10 | $22.23 \pm 2.44$ | **23.40** | 24.51 | 21.33 | 18.82 |
| Compositional Soft Attention | 81.32 | $21.08 \pm 2.02$ | 23.19 | **25.21** | 18.78 | **19.67** |

Table 5: Full results for FCE GED dataset. Similarly to BEA 2019, we note that the Weighted Soft Attention performs best on the token-level, as evaluated by $F_1$ score. However, that is mainly because the model assigns 1 to most tokens, causing recall to be high. This is evident by the significantly lower $F_{0.5}$ metric. Our Ranked Soft Attention and Compositional Soft Attention models achieve similar performance, significantly better than Weighted Soft Attention if evaluated on the $F_{0.5}$ and *MAP* metrics.

| | Doc $F_1$ | $F_1$ | $F_{0.5}$ | $P$ | $R$ | MAP |
|---|---|---|---|---|---|---|
| Random Uniform | - | $5.46 \pm 0.18$ | 4.27 | 3.72 | 10.24 | 8.46 |
| Longformer Weighted Soft Attention | 93.41 | $6.89 \pm 0.19$ | 4.43 | 3.58 | **92.39** | 7.81 |
| Longformer Soft Attention Top-K | **94.42** | $5.90 \pm 2.21$ | 5.19 | 4.80 | 7.67 | 7.85 |
| Longformer Ranked Soft Attention | 92.63 | $14.13 \pm 2.27$ | 11.92 | 10.80 | 20.54 | 11.97 |
| Compositional Soft Attention | 91.85 | **25.46** $\pm 0.97$ | **20.12** | **17.66** | 45.85 | **26.82** |

Table 6: Full results for the IMDb-Pos Sentiment Detection dataset. We note that our Compositional Soft Attention architecture performs significantly better across all token-level metrics.

| | Doc $F_1$ | $F_1$ | $F_{0.5}$ | $P$ | $R$ | MAP |
|---|---|---|---|---|---|---|
| Random Uniform | - | $6.02 \pm 0.23$ | 4.80 | 4.23 | 10.46 | 9.90 |
| Longformer Weighted Soft Attention | 93.52 | $8.34 \pm 0.98$ | 5.40 | 4.37 | **93.38** | 10.59 |
| Longformer Soft Attention Top-K | **94.60** | $5.71 \pm 3.63$ | 5.02 | 4.65 | 7.40 | 9.03 |
| Longformer Ranked Soft Attention | 93.92 | $19.62 \pm 0.61$ | 16.98 | 15.63 | 27.27 | 16.44 |
| Compositional Soft Attention | 90.82 | **27.27** $\pm 3.68$ | **22.70** | **20.42** | 41.22 | **29.78** |

Table 7: Full results for the IMDb-Neg Sentiment Detection dataset. Similarly to IMDb-Pos, we note that our Compositional Soft Attention architecture performs significantly better across all token-level metrics.

| | W-SA | R-SA | C-SA |
|---|---|---|---|
| We | 0.99 | 0.00 | 0.07 |
| have | 0.99 | 0.01 | 0.23 |
| our | 0.99 | 0.00 | 0.20 |
| knowledge | 0.99 | 0.85 | 0.99 |
| easier | 0.99 | 0.96 | 0.87 |
| and | 0.99 | 0.00 | 0.80 |
| we | 0.99 | 0.06 | 0.59 |
| learn | 0.99 | 0.02 | 0.33 |
| more | 0.98 | 0.10 | 0.98 |
| easier | 0.61 | 0.78 | 0.92 |
| too | 0.99 | 0.81 | 0.00 |
| . | 0.85 | 0.01 | 0.48 |

(a) Sample token-level predictions for BEA 2019 positive sample (beginner learner). Compositional Soft Attention finds all evidence, but also scores neighboring tokens highly. Ranked Soft Attention on the other hand attends to less neighbouring tokens. We note that this might explain the performance differences between Grammatical Error Detection and Sentiment Detection datasets, as the the former annotations are more concentrated than the latter.

| | W-SA | R-SA | C-SA |
|---|---|---|---|
| You | 0.99 | 0.05 | 0.02 |
| said | 0.99 | 0.86 | 0.05 |
| that | 0.99 | 0.84 | 0.04 |
| it | 0.99 | 0.97 | 0.13 |
| was | 0.99 | 0.99 | 0.05 |
| a | 0.99 | 0.99 | 0.10 |
| perfect | 0.08 | 0.99 | 0.00 |
| evening | 0.66 | 0.49 | 0.05 |
| out | 0.99 | 0.00 | 0.18 |
| but | 0.99 | 0.13 | 0.00 |
| it | 0.58 | 0.98 | 0.17 |
| was | 0.97 | 0.00 | 0.57 |
| n't | 0.99 | 0.76 | 0.00 |
| like | 0.16 | 0.68 | 0.01 |
| that | 0.99 | 0.86 | 0.44 |
| . | 0.99 | 0.07 | 0.08 |

(b) Excerpt from an FCE positive document (beginner learner) without any grammatical errors in the sentence. We note that both Weighted Soft Attention and Ranked Soft Attention find a lot of false positives. On the other hand, Compositional Soft Attention correctly does not much rationale.

| | W-SA | R-SA | C-SA |
|---|---|---|---|
| So | 0.15 | 0.34 | 0.29 |
| , | 0.99 | 0.02 | 0.00 |
| I | 0.99 | 0.00 | 0.00 |
| hope | 0.91 | 0.76 | 0.00 |
| I | 0.99 | 0.91 | 0.02 |
| will | 0.99 | 0.99 | 0.00 |
| get | 0.99 | 0.37 | 0.00 |
| my | 0.99 | 0.99 | 0.00 |
| money | 0.98 | 0.98 | 0.00 |
| back | 0.62 | 0.00 | 0.04 |
| , | 0.99 | 0.98 | 0.00 |
| it | 0.99 | 0.99 | 0.00 |
| was | 0.99 | 0.99 | 0.12 |
| very | 0.85 | 0.99 | 0.00 |
| disappointing | 0.99 | 0.99 | 0.00 |
| evening | 0.99 | 0.99 | 0.00 |
| out | 0.99 | 0.99 | 0.00 |
| in | 0.99 | 0.96 | 0.00 |
| my | 1.00 | 0.03 | 0.00 |
| life | 0.37 | 0.00 | 0.03 |
| ! | 0.99 | 0.00 | 0.36 |

(c) Excerpt from a positive FCE document (beginner learner). This sentence includes grammatical errors. Ranked Soft Attention manages to pick up some manually annotated rationale, while Compositional Soft Attention fails. However, Ranked Soft Attention also finds many false positives.

| | W-SA | R-SA | C-SA |
|---|---|---|---|
| In | 0.99 | 0.00 | 0.70 |
| the | 0.99 | 0.00 | 0.72 |
| end | 0.11 | 0.01 | 0.51 |
| , | 0.91 | 0.00 | 0.40 |
| the | 0.99 | 0.00 | 0.52 |
| restaurant | 0.89 | 0.00 | 0.46 |
| was | 0.99 | 0.00 | 0.53 |
| closed | 0.97 | 0.00 | 0.14 |
| because | 0.99 | 0.01 | 0.27 |
| you | 0.75 | 0.18 | 0.28 |
| did | 0.04 | 0.00 | 0.26 |
| n't | 0.94 | 0.14 | 0.09 |
| have | 0.85 | 0.00 | 0.18 |
| enough | 0.27 | 0.00 | 0.28 |
| staffs | 0.99 | 0.00 | 0.27 |
| , | 0.15 | 0.00 | 0.56 |
| this | 0.99 | 0.00 | 0.39 |
| evening | 0.99 | 0.00 | 0.33 |
| ! | 0.98 | 0.00 | 0.30 |

(d) An excerpt from a FCE positive document (beginner learner), for a sentence with grammatical errors. Ranked Soft Attention fails to provide any rationale, while Compositional Soft Attention correctly finds *","* as rationale, but fails to attend to neighboring tokens.

Figure 2: Example predictions for Grammatical Error Detection datasets. **W-SA** corresponds to Weighted Soft Attention, **R-SA** to Ranked Soft Attention, while **C-SA** to Compositional Soft Attention. We highlight words that human annotators marked as rationale in orange, while also marking true positives in green and false positives as red.

| | W-SA | R-SA | C-SA |
|---|---|---|---|
| " | 0.99 | 0.00 | 0.00 |
| jack | 0.14 | 0.00 | 0.00 |
| frost | 0.99 | 0.00 | 0.16 |
| , | 0.13 | 0.00 | 0.00 |
| " | 0.99 | 0.97 | 0.99 |
| is | 0.99 | 0.00 | 0.99 |
| one | 0.99 | 0.00 | 0.99 |
| of | 0.96 | 0.00 | 0.99 |
| those | 0.99 | 0.93 | 0.98 |
| dumb | 0.99 | 0.00 | 0.01 |
| , | 0.99 | 0.07 | 0.87 |
| corny | 0.99 | 0.95 | 0.99 |
| concoctions | 0.96 | 0.00 | 0.99 |
| that | 0.99 | 0.40 | 0.95 |
| attempts | 0.99 | 0.00 | 0.50 |
| to | 0.99 | 0.00 | 0.00 |
| be | 0.99 | 0.00 | 0.04 |
| a | 0.99 | 0.00 | 0.19 |
| heartwarming | 0.99 | 0.00 | 0.01 |
| family | 0.99 | 0.05 | 0.89 |
| film | 0.91 | 0.00 | 0.38 |
| , | 0.97 | 0.99 | 0.99 |
| but | 0.99 | 0.00 | 0.99 |
| is | 0.99 | 0.01 | 0.99 |
| too | 0.99 | 0.98 | 0.99 |
| muddled | 0.99 | 0.19 | 0.99 |
| in | 0.99 | 0.00 | 0.99 |
| its | 0.95 | 0.00 | 0.96 |
| own | 1.00 | 0.51 | 0.98 |
| cliches | 0.99 | 0.95 | 0.94 |
| and | 0.99 | 0.74 | 0.84 |
| predictability | 0.97 | 0.00 | 0.99 |
| to | 0.98 | 0.19 | 0.91 |
| be | 0.04 | 0.02 | 0.93 |
| the | 0.00 | 0.00 | 0.00 |
| least | 0.00 | 0.00 | 0.36 |
| bit | 0.98 | 0.01 | 0.43 |
| touching | 0.99 | 0.00 | 0.69 |
| . | 0.95 | 0.00 | 0.00 |

(a) Sample predictions for a sample negative review in the IMDb-Neg dataset. We find that Weighted Soft Attention assigns similar scores to most tokens, while Ranked Soft Attention and Compositional Soft Attention manage to provide more fine-grained predictions. Compositional Soft Attention appears to recover spans of rationale better than Ranked Soft Attention.

| | W-SA | R-SA | C-SA |
|---|---|---|---|
| we | 1.00 | 0.02 | 0.63 |
| go | 0.99 | 0.00 | 0.65 |
| to | 0.99 | 0.00 | 0.61 |
| see | 1.00 | 0.00 | 0.63 |
| jet | 0.99 | 0.00 | 0.66 |
| li | 0.99 | 0.29 | 0.86 |
| movies | 1.00 | 0.41 | 0.79 |
| because | 0.99 | 0.37 | 0.69 |
| we | 0.99 | 0.00 | 0.82 |
| want | 0.99 | 0.00 | 0.88 |
| to | 0.99 | 0.00 | 0.68 |
| see | 1.00 | 0.00 | 0.64 |
| jet | 0.99 | 0.00 | 0.83 |
| li | 0.99 | 0.01 | 0.71 |
| kicking | 0.14 | 0.49 | 0.86 |
| a | 0.99 | 0.21 | 0.70 |
| lot | 0.65 | 0.05 | 0.80 |
| of | 0.99 | 0.86 | 0.80 |
| ass | 1.00 | 0.00 | 0.83 |
| from | 0.99 | 0.00 | 0.76 |
| side | 0.14 | 0.00 | 0.82 |
| to | 1.00 | 0.81 | 0.84 |
| side | 0.99 | 0.99 | 0.56 |
| , | 0.73 | 0.99 | 0.39 |
| and | 0.99 | 0.99 | 0.69 |
| this | 0.99 | 0.99 | 0.72 |
| film | 0.99 | 0.87 | 0.92 |
| delivers | 0.99 | 0.92 | 0.68 |
| gangbusters | 0.99 | 0.98 | 0.91 |
| on | 0.99 | 0.98 | 0.82 |
| that | 0.99 | 0.99 | 0.97 |
| front | 0.99 | 0.99 | 0.91 |
| . | 0.99 | 0.99 | 0.67 |

(b) Sample prediction for a positive movie review in the IMDb-Pos review. Here we note how Compositional Soft Attention learns the correct ranking of tokens, as evidenced by the higher scores for true positives than false positives. However, it still failed to optimize correctly for the classification threshold, leading numerous false positives. This problem is not present in the Ranked Soft Attention for this sample.

| | W-SA | R-SA | C-SA |
|---|---|---|---|
| it | 0.99 | 0.00 | 0.02 |
| is | 0.99 | 0.00 | 0.02 |
| one | 0.99 | 0.00 | 0.08 |
| of | 0.99 | 0.00 | 0.08 |
| the | 0.70 | 0.00 | 0.05 |
| most | 1.00 | 0.00 | 0.00 |
| ludicrously | 0.99 | 0.00 | 0.00 |
| conceived | 0.99 | 0.00 | 0.01 |
| efforts | 0.99 | 0.00 | 0.06 |
| in | 0.99 | 0.00 | 0.02 |
| recent | 0.99 | 0.00 | 0.03 |
| history | 0.99 | 0.00 | 0.14 |
| . | 0.99 | 0.00 | 0.02 |

(c) Sample predictions for a negative review in the IMDb-Pos dataset. The gold token-level labels are all 0, as there are no positive rationale in the negative review. Weighted Soft Attention still assigns scores close to 1 to most tokens, while both Ranked Soft Attention and Compositional Soft Attention learns not to attend to any tokens. This shows how the increased token-level supervision signal helps these architectures to learn to provide better token-level rationale.

Figure 3: Example predictions for Sentiment Detection IMDb datasets. **W-SA** corresponds to Weighted Soft Attention, **R-SA** to Ranked Soft Attention, while **C-SA** to Compositional Soft Attention. We highlight words that human annotators marked as rationale in orange, while also marking true positives in green and false positives as red.