Appendix A Model Hyperparameters

Similar to Jain and Wallace (2019) we use FastText pretrained embeddings (Joulin et al., 2016) for the SST and ADR datasets, Glove pretrained embeddings (Pennington et al., 2014) for the IMDB and AG News datasets, while we use Word2Vec (Mikolov et al., 2013) from Gensim (Řehůřek and Sojka, 2010) to train embeddings for MIMIC. All embeddings are of size d = 300. We also replace all numbers in text with a special symbol q and initialise the embeddings of unknown words randomly from a normal distribution, $\mathcal{N}(0, 1)$. The embeddings are not trained alongside the rest of the model.

We train the models using default Adam learning rate (1e-3) with 1e-4 weight decay, which adds an l_2 regulariser across all parameters. We use 64 dimensional hidden representations for one-layered bi-LSTM and bi-GRU encoders and 128 dimensional hidden representation for the MLP encoder following Jain and Wallace (2019). For the CNN we use 4 kernels of sizes [1, 3, 5, 7], each with 32 filters, giving a final contextual representation $\mathbf{h}_i$ of size $N = 128$, with ReLU activation function on the output of the filters, as per Jain and Wallace (2019).

For BERT we use the pre-trained version from Wolf et al. (2020) and fine-tune with a learning rate of $1e-5$ all BERT parameters except from the word embeddings, to simulate the scenario with the rest of the encoders, and $1e-4$ for the remainder of the parameters. We train our models three times using different random seeds and a batch size of 8 for BERT and 32 for the rest of the models.

For Conv-TaSc we apply a CNN with 15 channels over the scaled embedding $\mathbf{e}_i$ from Lin-TaSc, keeping a single stride and a 1-dimensional kernel. This way, we ensure that input words remain context-independent. We then sum over the filtered scaled embedding $\mathbf{e}_i^f$, to obtain the scores s_{x_i}. We have also experimented with filter sizes of [2, 10, 20 , 30, 50] individually and simultaneously.

For the MIMIC dataset we also attempted to use LongFormer (Beltagy et al., 2020), which is a BERT version that has the ability to accept and deal with longer sequences. However due to the increasing time to train and evaluate the model, this BERT variant was abandoned. Additionally we attempted to use Hierarchical BERT to deal with the longer sequences, however increases where not substantial and run times where similarly increased. Finally, contrary to the remainder of the datasets to deal with the long sequences of MIMIC we truncated the 256 first tokens and 256 last tokens, following the suggestions of Sun et al. (2019). We experimented with using the first and the last 512 tokens, but the head and tails truncation approach yielded the best performances.

Appendix B Additional parameters with TaSc variants

In Table 6 we present the additional parameters introduced by each variant, with Lin-TaSc requiring the lowest number of parameters and Feat-TaSc the most.

TaSc Mechanism	Additional Parameters
Lin-TaSc	$\lvert\mathbf{V}\rvert$
Feat-TaSc	$\lvert\mathbf{V}\rvert \times d$
Conv-TaSc	$\lvert\mathbf{V}\rvert + d \times n + n$

Table 6: Additional parameters resulting from the proposed TaSc mechanisms where $|V|$ is the vocabulary size, d the embedding dimension and n the number of channels in a CNN.

Appendix C Reproducibility Results

Computational infrastructure used: For the experiments above we used NVIDIA's TESLA V100 GPU.

Dataset description: We consider the following datasets for text classification following Wiegreffe and Pinter (2019) and Jain and Wallace (2019):

SST: *Stanford Sentiment Treebank* consists of sentences tagged with sentiment on a 5-point-scale from negative to postive (Socher et al., 2013). Jain and Wallace (2019) removed sentences with neutral sentiment and labelled the remaining sentences to negative and positive if they have a score lower or higher than 3 respectively.

IMDB: The *Large Movie Reviews Corpus* consists of 50,000 movie reviews labelled either as positive or negative (Maas et al., 2011). We filter the dataset as per Jain and Wallace (2019) to include movie reviews with sequence length less than 400 words.

ADR: A dataset of ∼20,000 tweets with labels indicating whether a Twitter post contains an adverse drug reaction or not (Sarker et al., 2015).

Data-set	Enc()	No-TaSc		Lin-TaSc		Feat-TaSc		Conv-TaSc	
		Dot	Tanh	Dot	Tanh	Dot	Tanh	Dot	Tanh
SST	BERT	.89	.90	.90	.87	.87	.87	.90	.90
	LSTM	.77	.78	.77	.78	.77	.80	.79	.80
	GRU	.78	.78	.78	.79	.78	.79	.78	.79
	MLP	.75	.77	.78	.78	.80	.80	.79	.81
	CNN	.77	.77	.79	.80	.80	.79	.79	.78
ADR	BERT	.81	.81	.81	.79	.80	.80	.80	.81
	LSTM	.74	.75	.77	.76	.77	.77	.78	.76
	GRU	.76	.75	.77	.77	.76	.79	.77	.77
	MLP	.73	.78	.76	.76	.78	.77	.76	.76
	CNN	.74	.73	.77	.76	.77	.77	.78	.78
IMDB	BERT	.92	.92	.93	.92	.92	.92	.92	.92
	LSTM	.90	.89	.89	.89	.89	.89	.89	.89
	GRU	.90	.90	.89	.90	.89	.90	.89	.89
	MLP	.88	.88	.88	.88	.89	.88	.89	.88
	CNN	.89	.89	.90	.89	.89	.89	.89	.89
AG	BERT	.95	.95	.94	.94	.95	.95	.94	.95
	LSTM	.93	.93	.92	.93	.93	.93	.93	.93
	GRU	.93	.93	.93	.93	.93	.93	.93	.93
	MLP	.93	.93	.93	.92	.93	.92	.93	.93
	CNN	.93	.93	.93	.93	.93	.93	.93	.93
MIMIC	BERT	.84	.83	.85	.84	.86	.84	.85	.83
	LSTM	.88	.89	.89	.89	.89	.90	.90	.90
	GRU	.89	.90	.89	.89	.90	.90	.90	.90
	MLP	.90	.89	.88	.88	.89	.88	.89	.89
	CNN	.90	.89	.90	.90	.90	.90	.89	.90

Table 7: Validation set F1-macro average scores (3 runs) across datasets, encoders and attention mechanisms for models with and without TaSc (No-TaSc). Standard deviations do not exceed 0.01.

AG: A subset of the original news articles[13] dataset compiled by Jain and Wallace (2019) for topic categorisation (*Business* and *World* news).

MIMIC: A sample of discharge summaries from the MIMIC III dataset of health records (Johnson et al., 2016). The task is to recognise if a given summary has been labelled as relevant to acute or chronic anemia (Jain and Wallace, 2019).

Validation set predictive performances: In Table 7 we present predictive performances on the validation checks for reproducibility on models with TaSc and models without (No-TaSc).

Appendix D Detailed Experiment Results

In Tables 8, 9, 10 and 11 we present results in numbers for Figures 1 and 2. Tables 8 and 9 show the percentage of decision flips recorded by removing the highest scored token, across encoders and datasets respectively and compliment Figure 1. Tables 10 and 11 show the average fraction of tokens required to be masked in order to cause a change in prediction (decision flip) and compliment Figure 2.

[13] https://di.unipi.it/~gulli/AG_corpus_of_news_articles.html. Accessed on Sep 2019

	Enc()	No-TaSc	Lin-TaSc	Feat-TaSc	Conv-TaSc
α	**BERT**	4.8	6.2 (1.3)	**6.6** (1.4)	3.7 (0.8)
	LSTM	**6.2**	4.8 (0.8)	5.1 (0.8)	4.8 (0.8)
	GRU	**6.2**	5.7 (0.9)	5.5 (0.9)	5.4 (0.9)
	MLP	**8.0**	6.0 (0.8)	5.2 (0.7)	5.6 (0.7)
	CNN	**9.3**	6.2 (0.7)	5.7 (0.6)	5.4 (0.6)
$\nabla\alpha$	**BERT**	5.2	6.4 (1.2)	**7.4** (1.4)	3.6 (0.7)
	LSTM	6.5	10.9 (1.7)	**12.5** (1.9)	12.0 (1.9)
	GRU	6.3	11.3 (1.8)	**12.4** (2.0)	11.8 (1.9)
	MLP	9.8	12.0 (1.2)	**13.1** (1.3)	13.0 (1.3)
	CNN	10.1	12.0 (1.2)	**13.2** (1.3)	13.4 (1.3)
$\alpha\nabla\alpha$	**BERT**	5.7	8.0 (1.4)	**9.3** (1.6)	4.0 (0.7)
	LSTM	8.3	12.8 (1.5)	**13.6** (1.6)	13.1 (1.6)
	GRU	8.3	**14.2** (1.7)	13.9 (1.7)	13.4 (1.6)
	MLP	13.7	**14.6** (1.1)	13.9 (1.0)	13.8 (1.0)
	CNN	13.9	**14.7** (1.1)	14.5 (1.0)	14.6 (1.0)

Table 8: Mean average *percentage of decision flips* occurred by removing the most informative token, using the three TaSc variants and No-TaSc across encoders (higher is better).

	Dataset	No-TaSc	Lin-TaSc	Feat-TaSc	Conv-TaSc
α	**SST**	**16.7**	12.5 (0.7)	11.2 (0.7)	11.5 (0.7)
	ADR	4.5	**5.8** (1.3)	5.4 (1.2)	3.9 (0.9)
	IMDB	**6.6**	5.0 (0.8)	5.2 (0.8)	3.9 (0.6)
	AG	3.5	3.1 (0.9)	**3.8** (1.1)	2.7 (0.8)
	MIMIC	**3.0**	2.5 (0.8)	2.5 (0.8)	2.5 (0.8)
$\nabla\alpha$	**SST**	18.8	25.8 (1.4)	**27.5** (1.5)	25.4 (1.3)
	ADR	5.2	9.4 (1.8)	**10.6** (2.0)	8.1 (1.5)
	IMDB	7.1	7.6 (1.1)	9.2 (1.3)	**9.3** (1.3)
	AG	4.3	5.1 (1.2)	6.4 (1.5)	6.4 (1.5)
	MIMIC	2.5	**4.8** (1.9)	**4.8** (1.9)	**4.8** (1.9)
$\alpha\nabla\alpha$	**SST**	24.1	29.4 (1.2)	**29.5** (1.2)	27.2 (1.1)
	ADR	6.0	10.3 (1.7)	**11.1** (1.8)	8.2 (1.4)
	IMDB	10.3	**12.0** (1.2)	11.0 (1.1)	10.8 (1.0)
	AG	5.2	6.4 (1.2)	**8.0** (1.5)	6.9 (1.3)
	MIMIC	4.3	**6.2** (1.5)	5.7 (1.3)	5.8 (1.3)

Table 9: Mean average *percentage of decision flips* occurred by removing the most informative token, using the three TaSc variants and No-TaSc across datasets (higher is better).

	Enc()	No-TaSc	Lin-TaSc	Feat-TaSc	Conv-TaSc
α	**BERT**	.59	.46 (0.8)	**.44** (0.7)	.56 (0.9)
	LSTM	.56	**.48** (0.9)	.51 (0.9)	.52 (0.9)
	GRU	.57	**.45** (0.8)	.49 (0.9)	.50 (0.9)
	MLP	**.41**	.43 (1.0)	.44 (1.1)	.46 (1.1)
	CNN	.45	.47 (1.0)	.47 (1.0)	**.44** (1.0)
$\nabla\alpha$	**BERT**	.52	.34 (0.6)	**.31** (0.6)	.58 (1.1)
	LSTM	.44	.21 (0.5)	**.17** (0.4)	.19 (0.4)
	GRU	.46	**.17** (0.4)	.18 (0.4)	.19 (0.4)
	MLP	.21	.18 (0.8)	**.17** (0.8)	**.17** (0.8)
	CNN	.30	**.17** (0.6)	**.17** (0.6)	**.17** (0.6)
$\alpha\nabla\alpha$	**BERT**	.52	**.29** (0.6)	**.29** (0.6)	.57 (1.1)
	LSTM	.44	.20 (0.5)	**.16** (0.4)	.18 (0.4)
	GRU	.43	**.16** (0.4)	.17 (0.4)	.18 (0.4)
	MLP	.17	**.15** (0.9)	.16 (0.9)	.16 (0.9)
	CNN	.27	**.16** (0.6)	**.16**(0.6)	**.16** (0.6)

Table 10: Mean average *fraction of tokens* required to cause a decision flip, using the three TaSc variants and No-TaSc across encoders (lower is better).

	Dataset	No-TaSc	Lin-TaSc	Feat-TaSc	Conv-TaSc
α	**SST**	**.45**	.49 (1.1)	.48 (1.1)	.50 (1.1)
	ADR	.88	**.72** (0.8)	.76 (0.9)	.77 (0.9)
	IMDB	.38	**.31** (0.8)	.36 (1.0)	.41 (1.1)
	AG	.59	.52 (0.9)	**.49** (0.8)	.54 (0.9)
	MIMIC	.28	**.24** (0.9)	.26 (0.9)	.26 (0.9)
$\nabla\alpha$	**SST**	.35	.24 (0.7)	**.20** (0.6)	.27 (0.8)
	ADR	.78	.38 (0.5)	**.36** (0.5)	.47 (0.6)
	IMDB	.19	**.10** (0.5)	.12 (0.6)	.14 (0.7)
	AG	.49	.35 (0.7)	**.28** (0.6)	.35 (0.7)
	MIMIC	.13	**.02** (0.2)	.03 (0.3)	.07 (0.5)
$\alpha\nabla\alpha$	**SST**	.33	.23 (0.7)	**.19** (0.6)	.26 (0.8)
	ADR	.77	**.34** (0.5)	.35 (0.5)	.47 (0.6)
	IMDB	.17	**.07** (0.4)	.10 (0.6)	.13 (0.7)
	AG	.46	.31 (0.7)	**.25** (0.5)	.34 (0.7)
	MIMIC	.10	**.02** (0.2)	.03 (0.3)	.06 (0.7)

Table 11: Mean average *fraction of tokens* required to cause a decision flip, using the three TaSc variants and No-TaSc across datasets (lower is better).

Appendix E Comparing TaSc with Non-attention Input Importance Metrics

		Tanh							Dot						
		Non - TaSc			Lin-TaSc				Non-TaSc			Lin-TaSc			
Data	**Enc()**	**WO**	**x∇x**	**IG**	**WO**	**x∇x**	**IG**	$\alpha\nabla\alpha$	**WO**	**x∇x**	**IG**	**WO**	**x∇x**	**IG**	$\alpha\nabla\alpha$
SST	BERT	.29	.64	.51	.37	.30	.25	**.22**	**.32**	.62	.49	.35	.57	.51	.55
	LSTM	.25	.24	.20	.26	.33	**.19**	**.19**	.21	.23	**.19**	.21	**.19**	**.19**	**.19**
	GRU	.24	.22	.19	.29	.24	.20	**.18**	.24	.25	.23	.21	**.19**	**.19**	**.19**
	MLP	.36	.26	.24	.26	.20	.19	**.18**	.22	.19	**.18**	.24	.19	.19	**.18**
	CNN	.30	.25	.20	.27	.22	.20	**.19**	.22	.20	**.18**	.21	.20	.20	.19
ADR	BERT	.83	.91	.89	.73	.55	.38	**.31**	.81	.90	.87	.68	.58	.52	**.50**
	LSTM	.82	.81	.80	.54	.42	.34	**.32**	.87	.88	.87	.42	.35	**.34**	**.34**
	GRU	.84	.84	.84	.49	.38	.36	**.35**	.79	.80	.80	.50	.40	.44	**.38**
	MLP	.71	.63	.57	.60	.36	.43	**.31**	.49	.43	**.39**	.49	.40	.44	.40
	CNN	.80	.78	.78	.57	.46	.39	**.37**	.77	.74	.74	.52	.43	.38	**.36**
IMDB	BERT	.23	.69	.43	.27	.14	.16	**.07**	.24	.72	.49	.26	.27	**.17**	.20
	LSTM	.18	.12	.07	.11	.13	.05	**.04**	.26	.09	.07	.07	.06	.06	**.05**
	GRU	.18	.12	.07	.11	.06	.05	**.04**	.27	.15	.08	.09	**.05**	**.05**	**.05**
	MLP	.16	**.05**	**.05**	.07	**.05**	**.05**	**.05**	.18	.07	.06	.09	**.05**	**.05**	**.05**
	CNN	.21	.09	.07	.18	.07	.06	**.05**	.27	.07	.06	.14	.07	.06	**.05**
AG	BERT	.62	.78	.56	.64	.58	.54	**.50**	.56	.76	.60	.56	.59	**.55**	.60
	LSTM	.53	.51	**.30**	.47	.37	.31	.38	.47	.52	**.35**	.43	.40	.36	.46
	GRU	.45	.36	.31	.50	.30	.24	**.20**	.54	.40	.30	.36	.24	.23	**.22**
	MLP	.53	.24	.25	.53	.23	.23	**.19**	.44	.25	.23	.40	**.19**	.25	**.19**
	CNN	.55	.38	.28	.48	.29	.24	**.20**	.53	.35	.25	.39	.27	.23	**.21**
MIMIC	BERT	.24	.67	.43	.31	.10	.04	**.03**	.21	.57	.26	.25	.07	**.05**	**.05**
	LSTM	.35	.32	.12	.37	**.01**	.02	**.01**	.28	.40	.30	.40	**.01**	.02	**.01**
	GRU	.20	.24	.23	.46	**.01**	.02	**.01**	.36	.18	.08	.42	**.01**	.02	**.01**
	MLP	.40	.03	.22	.18	**.01**	.02	**.01**	.13	.04	.03	.16	**.02**	**.02**	**.02**
	CNN	.26	.15	.02	.52	**.01**	**.01**	**.01**	.43	.09	**.02**	.49	.03	**.02**	**.02**

Table 12: Average fraction of tokens required to cause a decision flip using the best performing attention-based ranking ($\alpha\nabla\alpha$) with TaSc, *Word omission*, (**WO**), *InputXGrad*, ($\nabla\mathbf{x}$) and *Integrated Gradients* (**IG**). Underlined values denote that Lin-TaSc is better and **bold** values denote the best performing method row-wise. (lower is better)