BERT, but Better: Improving Robustness using Human Insights

Michael Pieke – 8474752

Master's Thesis Artificial Intelligence

Supervised by: Dong Nguyen, Elize Herrewijnen, Yupei Du
Second reader: Albert Gatt

July 17, 2023

Contents

List of Figures

List of Tables

Glossary

Acronyms/Abbreviations

1. **LLM**: Large Language Model.
2. **AI**: Artificial Intelligence.
3. **NLP**: Natural Language Processing.
4. **ML**: Machine Learning.
5. **OOD**: Out-of-distribution.
6. **ID**: In-distribution.
7. **Rationale**: Annotator Rationale.
8. $\boldsymbol{L_{labels}}$: Cross-entropy loss between predicted and actual labels.
9. $\boldsymbol{L_{att}}$: Additional loss term that calulates the difference between a model's attention scores and rationales for each token in an input sequence.
10. **CE + R**: Cross-Entropy + Rationales. A pre-trained BERT classifier fine-tuned using the combined loss of L_{labels} and L_{att}, namely the cross-entropy loss of the predicted and actual labels as well as the loss between the model's attention scores for individual tokens in the input sequence and the rationales corresponding to those tokens.
11. **BERT-base**: A 'vanilla' pre-trained BERT classifier that uses only L_{labels} during fine-tuning.
12. **Only Rationales**: A pre-trained BERT classifier fine-tuned with L_{labels}, but all words in the input sequence that are not rationales are masked.
13. **CE + F**: Cross-Entropy + Flipped Rationales. A pre-trained BERT classifier fine-tuned with both L_{labels} and L_{att}. However, in this case, the rationale vectors used to compute L_{att} are 'flipped', meaning all zeros are converted to ones and all ones are converted to zeros.

Additional Defintions

1. **Annotator Rationale**: A vector representation of human insights on which portions of a text are most indicative of the meaning of that. These insights are gained by instructing humans to highlight these portions.
2. **Rationale augmentation**: the process of including annotator rationales as part of the training signal of a machine learning model.
3. **Rationale-augmented model**: A model to which rationale augmentation has been applied.
4. **Attention regularisation**: A specific form of rationale augmentation that includes a loss term, such as L_{att}, that accounts for the difference between a model's attention scores for an input sequence and the corresponding rationales for that sequence.

Abstract

Pre-trained transformers are highly effective across numerous Natural Language Processing (NLP) tasks, yet their ability to generalise to new domains remains a concern due to their tendency to rely on spurious correlations. Consequently, this thesis investigates the impact of token-level human supervision to enhance BERT's generalisation capabilities. Although the benefits of token-level insights have been shown to improve the performance of these models, few studies have examined the effect of these insights to improve generalisation. Consequently, this work explores the potential of human supervision to guide BERT's attention mechanism towards salient features, thereby improving generalisation across domains. Results from experiments in both binary and multi-label classification scenarios demonstrate not only substantial gains in out-of-distribution (OOD) performance in few-shot contexts, but also a closer alignment between the model's attention scores and salient features identified by human annotators. By emphasising the role of human insight in transformer models, this thesis contributes to the ongoing discourse on enhancing performance and explainability in NLP applications.

1 Introduction

1.1 The issue of spurious correlations

Due to recent advancements in deep learning, platforms powered by Large Language Models (LLMs), such as OpenAI's ChatGPT, are often seen by the public as synonymous with the concept of Artificial Intelligence (AI). These models are able to achieve impressive, often above human-level performance on various Natural Language Processing (NLP) tasks due to their ability to comprehend highly complex patterns in textual data, making them the go-to approach for both academic and industrial purposes (Min et al., 2021). Having been pre-trained on a vast array of unlabeled documents, these models are able to gain extensive knowledge about complex linguistic patterns, which can then be applied to execute downstream tasks with high success (Qiu et al., 2020). Given this preconceived knowledge, such models have been shown to be much more robust to novel data than previous NLP approaches (Tu, Lalwani, Gella, & He, 2020; Hendrycks et al., 2020).

Despite the efficacy of such models, recent work has shown that LLMs still often rely on invalid heuristics derived from spurious correlations present in their training data to be effective at the task they are meant to perform (McCoy, Pavlick, & Linzen, 2019; Liusie, Raina, Raina, & Gales, 2022; Gururangan et al., 2018; Poliak, Naradowsky, Haldar, Rudinger, & Van Durme, 2018; Nie et al., 2020). For example, Liusie et al. (2022) observe a substantial decrease in accuracy on a sentiment analysis task when a BERT model trained on IMDB movie reviews is evaluated on tweets. They attribute this decline in performance to the model having learned that stopwords, such as "and", "or" and "but", were distributed differently throughout documents with a positive sentiment compared to those with a negative sentiment. While using this difference as a heuristic to distinguish between classes was effective on the movie reviews the model was trained on, this heuristic is no longer applicable to the tweet dataset. In other words, the distribution of stopwords should not be relied upon as an indicator of the sentiment of a document.

This example demonstrates that LLMs do not necessarily always encode the features required to correctly perform a task. Since LLMs are often evaluated on a dataset using a standard train-test split (Hupkes et al., 2022), researchers may not realise that the model is not functioning as intended, as the test set is considered in-distribution (ID), namely sampled from the same distribution as the training set. Consequently, erroneous heuristics derived from the training data can successfully be applied to the test set (McCoy et al., 2019). However, a model that relies on correlationally, but not causally, related features does not capture the core elements of the task it is designed to perform (G. Marcus, 2018). Therefore, when presented with out-of-distribution (OOD) data, namely data sampled from a different distribution than the training data, such a model may encounter difficulties because this novel data is unlikely to contain the same spurious correlations present in the training data (Arjovsky, 2020). To make these models applicable to diverse datasets, they must be able to *generalise*.

1.2 What is generalisation and why is it important?

Generalisation can be broadly defined as "*the ability to successfully transfer representations, knowledge, and strategies from past experience to new experiences*" (Hupkes et al., 2022). Models that can generalise are developed for a range of purposes (Hupkes et al., 2022). For instance, such models can be used to advance linguistic theories and deepen our understanding of how humans encode and represent their knowledge of language (Baroni, 2022; G. F. Marcus, 1999). Others aim to create fair and inclusive ML models that do not rely on sensitive attributes such as race or gender (Bolukbasi, Chang, Zou, Saligrama, & Kalai, 2016; Bordia & Bowman, 2019; Ravfogel, Elazar, Gonen, Twiton, & Goldberg, 2020). While these motivations are valid, this thesis focuses on creating models that can support decision making in practical scenarios involving data from diverse sources.

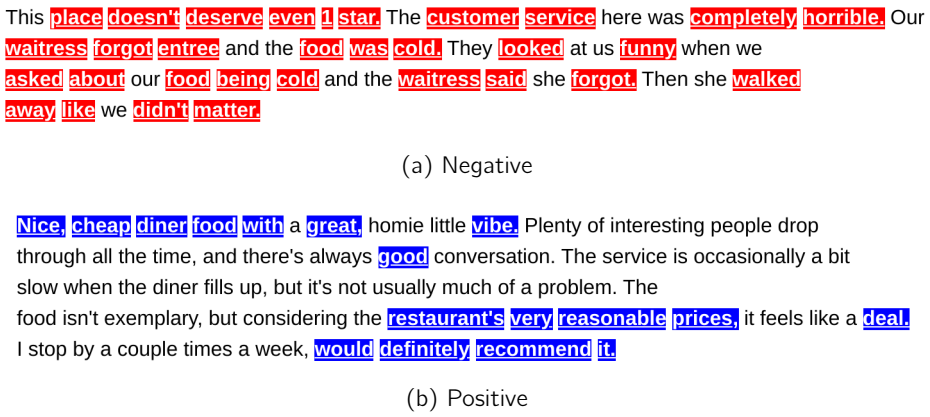
Relying too heavily on potentially flawed AI systems to make important decisions can have negative consequences in sectors such as finance (Ziora, 2016; Hair Jr & Sarstedt, 2021; Canhoto & Padmanabhan, 2015), healthcare (Yu, Beam, & Kohane, 2018; S. H. Park & Han, 2018; Rajpurkar, Chen, Banerjee, & Topol, 2022), and law (Betts & Jaep, 2016; Dale, 2019). For example, automated sentiment analysis can help companies analyse trends and gain a competitive advantage (Ziora, 2016). However, using a model that relies on unrepresentative or incomplete data can lead to poor investments and bankruptcy (Hair Jr & Sarstedt, 2021; Canhoto & Padmanabhan, 2015). Moreover, the healthcare sector is using ML systems for clinical diagnoses, but symptoms

and conditions that are underrepresented in the training data can lead to dangerous misdiagnoses (Yu et al., 2018; S. H. Park & Han, 2018). Furthermore, it may be the case that labeled data is scarce in certain practical situations. These examples highlight the importance of empirically testing whether these models can be relied upon or whether additional measures must be taken to ensure they are robust to the difficulties which practical settings present.

1.3 Using annotator rationales to increase generalisability

Numerous approaches have been investigated to enhance the generalisation capabilities of LLMs, including data augmentation and ensemble learning (Z. Shen et al., 2021). While these methods have proven effective at enhancing OOD performance, an entirely different method is gradually emerging that modifies the traditional supervised learning approach by incorporating token-level insights from humans. Humans have shown better generalisation capabilities across domains than machine learning models (Geirhos et al., 2018), potentially due to the human ability to gain fundamental insights about a task that can be applied across various domains (Simon, 1990). Therefore, integrating these insights during training may help enhance the generalisation capabilities of machine learning models. Many researchers have done so using so-called 'annotator rationales' to train models to pay attention to the same features humans do (D. Zhang et al., 2021; Kanchinadam, Westpfahl, You, & Fung, 2020; Hartmann & Sonntag, 2022; O. Zaidan & Eisner, 2008; Pruthi et al., 2022; Mathew et al., 2021; O. Zaidan, Eisner, & Piatko, 2007; Wang, Sharma, & Bilgic, 2022; Carton, Kanoria, & Tan, 2022; Bao, Chang, Yu, & Barzilay, 2018; Melamud, Bornea, & Barker, 2019; Stacey, Belinkov, & Rei, 2022; Y. Zhang, Marshall, & Wallace, 2016; Zou et al., 2021).

These annotator rationales, or rationales for short, are created by instructing humans to highlight words or sentences that are indicative of the class label of a document. For example, Figure 1 displays two movie reviews with corresponding rationales for a sentiment analysis task. As noted by the human annotators, words such as "nice" and "great" indicate a positive sentiment, while words such as "completely horrible" point to a negative sentiment. Models using these rationales as additional information, which I call *rationale-augmented models*, have seen varying degrees of improvements compared to models that have not been provided these annotations. However, for most of these works, model performance has been measured using ID test sets (Pruthi et al., 2022; Kanchinadam et al., 2020; O. F. Zaidan, Eisner, & Piatko, 2008; Mathew et al., 2021; O. Zaidan et al., 2007; Wang et al., 2022; Carton et al., 2022) meaning these observed performances are not guaranteed to carry over to new domains. However, when a model is evaluated on an OOD dataset, the underlying task remains the same, despite the specific instances it encounters being new and unfamiliar . Therefore, a model that has been trained to focus on the fundamental, task-related cues rather than spurious correlations may be more adaptable and thus perform better on the OOD data. As such, this thesis assesses whether using these rationales as part of the training signal during fine-tuning, a process I call *rationale augmentation*, can also enhance an LLM's ability to identify and leverage salient cues in OOD samples, thereby improving its ability to generalise.



(a) Negative

(b) Positive

Figure 1: Two restaurant review taken from from Yelp.com with a negative (top) and positive (bottom) sentiment. Annotator rationales are respresented by the highlighted and underlined words.

1.4 Extending current literature on the OOD performance of rationale-augmented models

This thesis unifies valuable elements of previous works to further understand the practical utility of rationale-augmented models in OOD environments. While the field of research into the OOD performance of these models is still in its infancy, initial experiments have shown promising results. Bao et al. (2018) trained a rationale-augmented LSTM that was able to successfully transfer the knowledge it had obtained from the beer reviews it was trained on to an OOD dataset of hotel reviews. Stacey et al. (2022) observed similar results and went a step further to understand how rationale augmentation affects the internals of the BERT model they trained. Analyses of the attention weights of both rationale augmented and baseline (i.e. not augmented with rationales) BERT models, suggested the former model to focus less on spurious features, such as stopwords, than the latter model. These insights serve as invaluable stepping stones towards a comprehensive understanding of the effects of rationale augmentation and how these effects can improve an LLM's ability to generalise.

We can also draw from other related studies that do not specifically address OOD performance. Notably, Pruthi et al. (2022) evaluated multiple rationale augmentation approaches on different tasks and training set sizes, revealing that the benefits of rationale augmentation vary depending on all three of these factors. For example, one rationale augmentation approach, coined *attention regularisation*, is shown to be especially beneficial for sentiment analysis with relatively few training instances. Although the generalisability of these models was not evaluated, this work provides key insights that can also be used to evaluate the OOD performance of rationale-augmented models from different perspectives. In summary, by leveraging insights from related studies, we can gain a more nuanced understanding of the generalisation capabilities of rationale-augmented models and create a more robust framework for evaluating their practical applicability.

1.5 Research question and contributions

To properly define a research question to investigate the effect of rationale augmentation on the OOD performance of ML models, there are two clarifications to be made. Firstly, researchers have explored many approaches to augment models with annotator rationales. While the verdict is still out on which of these approaches is the 'best', my experiments solely concern the attention regularisation approach mentioned in Section 1.3. This approach has been most widely used in previous works and shown to be effective for a variety of text classification tasks (Stacey et al., 2022; Pruthi et al., 2022; Bao et al., 2018; Mathew et al., 2021; Kanchinadam et al., 2020). Attention regularisation employs an additional loss term during training to account for the disparity between the model's attention scores for tokens in an input sequence and the corresponding annotator rationales. By incorporating this loss term, the approach aims to encourage the model to prioritise the words and phrases deemed significant by humans, while simultaneously maintaining its focus on making accurate predictions. Regarding the second clarification, I limit my experiments to pre-trained transformers, as current state of the art NLP models build upon such architectures (Lauriola, Lavelli, & Aiolli, 2022). Based on these clarifications, my main research question is as follows

RQ: How does attention regularisation affect the OOD performance of pre-trained transformers?

This research question was studied from three different perspectives. Firstly, most works evaluate their approach to rationale augmentation in binary classification settings, as annotator rationales are widely available for this task. However, no works have yet assessed the effects of rationale augmentation in scenarios with more than two class labels. As noted by Wang et al. (2022), these scenarios require additional considerations, such as a modified loss function that can incorporate rationales from multiple classes. To better understand the practical effectiveness of rationale augmentation, it is important to investigate its impact in different problem settings, such as binary and multi-label classification. This exploration can help assess whether rationale augmentation serves as a versatile tool applicable to a variety of NLP tasks or only benefits a handful of these tasks[1]. To this end, this study focuses on two distinct tasks: (1) sentiment analysis, a binary classification task, and (2) emotion detection, a multi-label classification task. This comparison serves as a case study, investigating whether the benefits of rationale augmentation extend beyond binary classification, offering initial insights into the effectiveness of such augmentation in more intricate scenarios. With these considerations in mind, my first sub-question can be defined as follows:

[1]Other researchers have experimented with different forms of human explanations, such as free-text and structured explanations (e.g. Ross, Peters, and Marasović (2022) and Yao, Chen, Ye, Jin, and Ren (2021)), to increase the OOD performance of their models. While these approaches are promising, they fall outside the scope of this thesis.

SQ1: Does attention regularisation affect the OOD performance of pre-trained transformers similarly for sentiment analysis and emotion detection?

Sentiment analysis is a binary classification task that has been widely used in previous works on rationale augmentation, making it a useful benchmark for comparing the performance of rationale-augmented models. On the other hand, emotion detection is a multi-label classification problem to which rationale augmentation has not yet been applied. By driving the body of literature on rationale augmentation forward to new domains and tasks while also remaining anchored in traditional approaches, addressing this research question provides a robust empirical basis for the effectiveness of rationale augmentation in different NLP applications.

To thoroughly investigate the impact of attention regularisation on pre-trained transformers, it is necessary to go beyond traditional evaluation measures such as accuracy and F1-score. To this end, this thesis provides valuable insights for practical applications by examining whether attention regularisation encourages the model to focus more on highlighted words or phrases. In doing so, this investigation deepens our understanding of the factors driving improved performance. Gaining insights into these factors can guide the development of new and improved approaches to enhance OOD performance, thereby advancing the field and improving the overall effectiveness of models in real-world applications. Moreover, these insights can facilitate the development of more trustworthy and explainable models, bolstering confidence in the reliability of these models beyond controlled laboratory settings, which is particularly important in critical domains such as healthcare. Based on these considerations, the following sub-question is addressed:

SQ2: To what extent does attention regularisation guide the attention mechanism of BERT to align with salient features highlighted by human annotators?

To answer this question, I follow previous studies, such as those by Sen, Hartvigsen, Yin, Kong, and Rundensteiner (2020), Herrewijnen, Nguyen, Mense, Bex, et al. (2021), and DeYoung et al. (2020), have evaluated the quality of their models by computing token-level feature importance scores using various methods and measuring the similarity of these scores to the annotator rationales of the same input sequence. In this thesis, these feature importances are based on the attention scores assigned to each token. However, there is some discussion about whether attention scores truly reflect the inner workings of these models. For example, Jain and Wallace (2019) show that explanations derived from attention mechanisms do not correlate well with other feature importance measures, while these other measures do correlate well with each other. While other explanation methods may be more reliable, I have specifically opted for attention-based explanations to understand whether incorporating attention scores in the loss function has a direct influence on the overall behaviour of the attention mechanism.

The final perspective addressed in this thesis relates to a model's ability to generalise from small amounts of data. As a lack of data is prevalent in many practical ML scenarios (Ravi & Larochelle, 2016), devising techniques to handle such environments can be valuable. For example, legal languages are highly specialised, meaning there are few people who can write, evaluate or label documents written in these languages (Sovrano, Palmirani, & Vitali, 2022). In these situations it is especially important to extract the most amount of useful information possible from the data that is available, such that unseen data can be handled effectively. Attention regularisation is a promising approach for such scenarios, as it has previously been shown to be particularly effective when less training data is available (Pruthi et al., 2022; Melamud et al., 2019; Bhat, Sordoni, & Mukherjee, 2021; Bao et al., 2018). Notably, when trained on 200 or less texts, an LSTM trained using attention regularisation by Bao et al. (2018) showed up to 5% improvements in accuracy on OOD data compared to the same model trained without rationales on a sentiment analysis task. Inspired by these findings, the final sub-question is formulated as follows:

SQ3: Is attention regularisation an effective method for reducing the amount of training data required to achieve a desirable level of OOD performance in pre-trained transformers?

While previous works have already addressed this question, no one aside from Bao et al. (2018) has done so in terms of OOD performance. Hence, this question provides two opportunities: (1) to understand whether the benefits observed by Bao et al. (2018) are also present in transformer-based models and (2) to find out whether attention regularisation provides similar benefits when training sets larger than 200 examples are used. To address this gap in the literature, this study aims to explore the consistency of findings on the impact of attention regularisation across different model architectures and training set sizes.

Answering these three questions allows novel insights to be gained into the practical utility of rationale-augmented models by expanding upon and unifying previous works, such as Bao et al. (2018), Stacey et al. (2022) and Pruthi et al. (2022). In turn, the results of this thesis can provide a more nuanced insight into the different effects that rationale augmentation can have on OOD performance and the considerations that one must make when creating AI systems which can be deployed in practice.

2 Literature Review

In this section, I review previous approaches to developing generalisable models, evaluate the success of these models, and examine the role of rationale-augmented models within this area of research. To provide more context for my research, Tables 1 and 2 outline the choices made by previous works regarding model architecture, tasks and datasets. Furthermore, to gain a deeper understanding of the challenges associated with developing generalisable models, I discuss the inherent difficulties in defining the concept of OOD in Section 2.1. This discussion provides a foundation for my dataset selection and approach to evaluate OOD performance. Next, I provide a brief overview of pre-trained transformers in Section 2.2 and discuss how their ability to generalise is limited. Previous works that address these limitations are discussed in Section 2.3, after which the two most prominent rationale augmentation approaches and their role in OOD performance are reviewed in Section 2.5. By examining these findings, I establish a basis for my experimental setup to address the research questions in Section 1.5.

Author	Task	Data
Kanchinadam et al. (2020)	Sentiment analysis, QA	IMDB, TREC QA, Insurance claims
Pruthi et al. (2022)	Sentiment Analysis, Question Answering	IMDB, NQ
Barrett, Bingel, Hollenstein, Rei, and Søgaard (2018)	Sentiment Analysis, Grammatical Errors, Hate Speech	Dundee, ZuCo
Mathew et al. (2021)	Hate Speech Detection	HateXplain
Wang et al. (2022)	Sentiment Analysis, Aviation Safety, (unclear)	IMDB, ASRS (no ref), AIvsCR
Carton et al. (2022)	Reading Comprehension, Fact Verification, NLI	MultiRC, FEVER, e-SNLI
Bao et al. (2018)	Sentiment Analysis	BeerAdvocate, TripAdvisor
Melamud et al. (2019)	Sentiment analysis	IMDB
Stacey et al. (2022)	NLI	SNLI, MLNI, HANS
Y. Zhang et al. (2016)	Reliability of biomedical journals, sentiment analysis	Risk of Bias, IMDB
Zou et al. (2021)	QA	PALRACE
A. Sharma, Miner, Atkins, and Althoff (2020)	Empathy Detection	TalkLife, Reddit

Table 1: An overview of related works that experiment with rationale augmentation for various types of neural network architectures. While by no means exhaustive, the reader can use this table to gain an intuition about the overall trends regarding rationale augmentation. Here, the specific tasks and datasets are provided. This overview is limited to extractive rationales, namely highlights of an input sequence deemed important by a human annotator. Many works experiment with other forms of rationales, such as natural language explanations. However, these works fall beyond the scope of this thesis.

Author	Model	Method	OOD?
Kanchinadam et al. (2020)	FFNN with attention	AR	No
Pruthi et al. (2022)	BERT	AR, MT	No
Barrett et al. (2018)	BiLSTM	AR	No
Mathew et al. (2021)	BERT, BiRNN	AR	No
Wang et al. (2022)	BERT	other	No
Carton et al. (2022)	BERT	other	No
Bao et al. (2018)	BiLSTM, CNN	AR	Yes
Melamud et al. (2019)	CNN, BERT	MT	No
Stacey et al. (2022)	BERT, DeBERTA	AR	Yes
Y. Zhang et al. (2016)	CNN	other	No
Zou et al. (2021)	BERT, RoBERTa, ALBERT	other	No
A. Sharma et al. (2020)	RoBERTa	MT	No

Table 2: An overview of the same works listed in Table 1. However, in this overview, the specific models and rational augmentation approaches used are enumerated. MT stands for 'multi-task learning' and AR stands for 'attention regularisation'

2.1 The challenges that distribution shifts pose to examining OOD performance

When evaluating a model's OOD performance, it is crucial to consider the types of distribution shifts that can occur between training and test data, as these shifts can fundamentally affect the model's performance on OOD samples. In Section 2.1.1, the main causes of distribution shift are discussed and how researchers can selectively induce and study these shifts. Section 2.1.2 examines the various methods used to induce these shifts and their implications for drawing meaningful conclusions from experiments.

2.1.1 Causes of distribution shift

Distribution, or dataset shifts occur when "*the testing (unseen) data experience a phenomenon that leads to a change in the distribution of a single feature, a combination of features, or the class boundaries.*" (Moreno-Torres, Raeder, Alaiz-Rodríguez, Chawla, & Herrera, 2012). If an ML model fails to learn patterns from its training data that account for and adapt to such phenomena, it will struggle to effectively handle the unseen data. Hence, distribution shifts frequently pose challenges for the deployment of ML models. Moreno-Torres et al. (2012) provide two main reasons why training and test distributions can differ systematically from one another. Firstly, *sample selection bias* may occur, where the process of collecting labelled data leads to a dataset that does not accurately represent the target population. For instance, when conducting sentiment analysis on social media posts for marketing research purposes, obtaining permission from post authors may be necessary to avoid privacy concerns. However, this approach can introduce sample selection bias if certain demographic groups are more likely to provide data (Rambocas & Pacheco, 2018). Secondly, the training data of a model may simply be unable to account for data produced by dynamic environments, despite being carefully collected. Such environments can play a vital role in NLP applications. For example, an effective chat-bot ought to successfully process input from many different users, even if it has not been specifically trained on data from these users (Shum, He, & Li, 2018)[2]. These scenarios can result in different types of distribution shifts, the most notable ones being *covariate shift*, where the input features of the test data are distributed differently from those of the training data, and *label/concept shift*, where the relationship between the data and its class labels differs between the training and test data.[3].

2.1.2 Dealing with distribution shifts

Inducing distribution shifts between the training and test data is important to assess a model's ability to generalise (Hendrycks et al., 2021) (See Figure 2 for an abstract example). Researchers can artificially induce such shifts by generating synthetic training or test sets (Y. Zhang, Baldridge, & He, 2019; McCoy et al., 2019; Bhargava, Drozd, & Rogers, 2021; Cui, Hershcovich, & Søgaard, 2022; Raunak, Kumar, & Metze, 2019). For example, Y. Zhang et al. (2019) constructed adversarial example datasets by using back translation and scrambling the word order to perturb texts extracted from Quora or Wikipedia. When state-of-the-art models, including BERT and bidirectional LSTMs, were trained on the unaltered versions of these texts, they performed remarkably poorly on the perturbed texts. Based on these results, the authors conclude that these adversarial examples indicate that the models are sensitive to the word order and syntactic structure of the input sequences as opposed to properly understanding the content of these sequences. In this scenario, the training data is naturally occurring, while the test data is artificial. To establish even more experimental control, one can generate both training and test sets synthetically (Hupkes, Dankers, Mul, & Bruni, 2020; Lake & Baroni, 2018). In this way, the researcher can specifically define how these two datasets differ from each other, providing even more certainty over the results of such an experiment. While such setups allow the experimenter to be confident in their interpretation of the results of such an experiment, the ecological validity of empirical results based on synthetic data is questionable, making such experiments less informative of practical settings.

Accordingly, other works have tested the OOD performance of LLMs given naturally occurring shifts between training and test data (Talman & Chatzikyriakidis, 2018; Lazaridou et al., 2021; Hendrycks et al., 2020; Desai & Durrett, 2020; Lu, Yang, Namee, & Zhang, 2022; Fisch et al., 2019; Miller, Krauth, Recht, & Schmidt, 2020; Clinchant, Jung, & Nikoulina, 2019). While some simply assume this shift takes place, others provide more

[2]A notable example of a successful chat-bot is OpenAI's ChatGPT, which is trained on conversational tasks and refined through feedback from users and workers. This enables ChatGPT to provide personalised recommendations and insights based on the user's preferences and interests, even when different users have distinct writing styles (Y. Shen et al., 2023).

[3]Multiple shifts may occur when experimenting with pre-trained models (Hupkes et al., 2022). This is because these models undergo three distinct processes, namely pre-training, fine-tuning and evaluation. This provides two separate opportunities for distribution shift to occur.

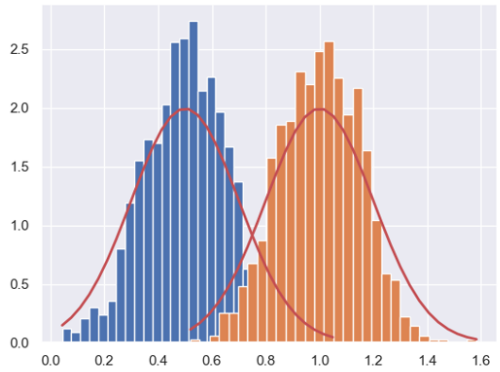


Figure 2: A hypothetical example of a distribution shift between two datasets. Suppose one dataset contains movie reviews while the other contains restaurant reviews. Then, the overlap between the two distributions encapsulates the features that these two types of reviews have in common. A model can thus be said to generalise effectively from one dataset to the other if it sufficiently captures this overlap.

justification for what constitutes as a naturally occurring shift. For example, Hendrycks et al. (2020) consider a dataset to be OOD if it either "*(1) contains metadata which allows us to naturally split the samples or (2) can be paired with a similar dataset from a distinct data generating process.*". While this approach better replicates real-world scenarios compared to using synthetic data, it becomes challenging to pinpoint the exact driving factors behind the distribution shift, making it less clear why the model succeeds or struggles in generalising. For instance, Fisch et al. (2019) curated the Machine Reading for Question Answering (MRQA) dataset by collecting examples from various sub-domains, with which they assess the OOD performance of models built on powerful architectures such as BERT and XLNet. While their study yielded valuable insights into which architectures are most robust to OOD data, Miller et al. (2020) pointed out that the data points collected did not just differ in terms of their source, but also in other ways, such as in the collection procedure, and crowd worker population. Despite efforts to address these factors, Miller et al. (2020) were still unable to pinpoint the exact reason for the decrease in OOD performance observed in their study. This example illustrates the intricate nature of real-world scenarios, which cannot be fully replicated through artificially induced shifts in data, a consideration that is taken seriously by many researchers in the field of deep learning as a whole (Taori et al., 2020; Liao, Taori, Raji, & Schmidt, 2021; Radford et al., 2021; Wortsman et al., 2022; Pham et al., 2021; Koh et al., 2021).

To understand the utility of attention regularisation in practical scenarios, my experiments involve naturally occurring shifts. Although naturally occurring distribution shifts between training and test data are less controlled and more ambiguous than artificially induced ones, this ambiguity reflects the complexity and diversity of real-world scenarios. Although this decision comes at the cost of experimental control, I consider these experiments to better assess a model's ability to generalise to new and diverse situations.

2.2 Pre-trained transformers

Pre-trained transformers have demonstrated impressive results in a range of NLP applications, and have also shown potential for handling naturally occurring distribution shifts (Hendrycks et al., 2020; Clinchant et al., 2019). Introduced by Vaswani et al. (2017), transformer-based architectures uses its attention mechanism to capture complex linguistic patterns by modelling a number of different relationships between words in a document. Due to the scarcity of labelled data for specific tasks, various approaches have been explored to first train transformer models to understand general patterns in unlabelled textual data via proxy tasks, such as predicting the next token or sentence of an input sequence. These approaches have given rise to well-known "pre-trained transformer models", such as the Generative Pre-trained Transformer (GPT) (Radford, Narasimhan, Salimans, Sutskever, et al., 2018) and Bidirectional Encoder Representations from Transformers (BERT) (Devlin, Chang, Lee, & Toutanova, 2019). Having been exposed to a vast amount of textual data in which many fundamental linguistic patterns can be found, such models no longer need to learn these patterns from scratch from task-specific, labelled data that may be in short supply. In doing so, these models can learn task or domain-specific features more effectively without overfitting to their training data (Qiu et al., 2020). Due

to the promising capabilities of these models, the following years have seen the further development of newer pre-trained transformer architectures, often yielding additional improvements over their predecessors (Sanh, Debut, Chaumond, & Wolf, 2019; Y. Liu et al., 2019; Yang et al., 2019; Lan et al., 2020; Z. Zhang, Liu, & Razavian, 2020; Beltagy, Lo, & Cohan, 2019).

2.3 The continued reliance of pre-trained transformers on spurious correlations

Due to the success of pre-trained transformers, it has since become the norm for many researchers to compare a number of these models when experimenting with the generalisability of these models (X. Liu et al., 2020; C. Zhu et al., 2020; Tu et al., 2020; Kavumba, Takahashi, & Oda, 2022; Hupkes et al., 2022; Hendrycks et al., 2020). Despite yielding significant improvements over previous state-of-the-art models, researchers have created so called "challenge" or "adversarial" datasets which have shown that pre-trained transformers are not immune to the problem of spurious correlations (Glockner, Shwartz, & Goldberg, 2018; McCoy et al., 2019; Y. Zhang et al., 2019; Poliak et al., 2018). For instance, McCoy et al. (2019) created the HANS dataset to test whether models use such heuristics, which contains entries that either support or contradict these heuristics. Consider the following example from McCoy et al. (2019): "The banker saw the actor" can be inferred from the premise "The banker near the judge saw the actor", whereas "The lawyer visited the doctors" cannot be inferred from "The doctors visited the lawyer," despite both premises and conclusions sharing five words. The HANS dataset provides examples where the overlap of words between premise and hypothesis is not always a reliable heuristic to determine the validity of an inference. Various models, including BERT, were found to have difficulty in correctly classifying examples in the HANS dataset where such a heuristic does not hold true. Similar results have been reported by other studies (e.g. Y. Zhang et al. (2019), Glockner et al. (2018) and Gururangan et al. (2018)), suggesting that these models may rely too heavily on superficial linguistic patterns when making inferences.

Despite the aforementioned difficulties, it is important to note that pre-trained transformers are still able to encode complex linguistic features, as evidenced by numerous studies (Hewitt & Manning, 2019; Bau et al., 2018; Tenney, Das, & Pavlick, 2019; Warstadt & Bowman, 2020). These models possess the capacity to capture the linguistic complexity required for the task at hand, but are less likely to do so when easily quantifiable spurious features are readily available in the training data (Lovering, Jha, Linzen, & Pavlick, 2021). The prevalence of spurious correlations and the preference for simpler features over complex ones has been noted in computer vision tasks as well (Nam, Cha, Ahn, Lee, & Shin, 2020; Le Bras et al., 2020), indicating that this issue is pervasive in the field of deep learning. To tackle this challenge, Lovering et al. (2021) advocate for the development of neural language models with explicit inductive biases that prioritise target features over spurious ones.

2.4 Increasing OOD performance through inductive bias

Various approaches have been explored to instil inductive biases into NLP models to increase the generalisation capabilities of these models. For example, data augmentation strategies strive to artificially introduce more variance into the training data by creating perturbations of existing samples (Bordia & Bowman, 2019; Ravfogel et al., 2020; Schick & Schütze, 2021; Wei & Zou, 2019). While these perturbations differ from their original counterparts in terms of superficial features, the core features are intended to remain the same. As such, this approach can be used to expand the training set of a model to contain more examples which contain relevant features. In this way, spurious features are less prominent in the training data, giving the model less incentive to focus on such features (Feng et al., 2021). Other methods take a different approach. For example, ensemble learning involves training multiple ML models on the same task, each of which is able to handle a certain subset of training instances. By merging the predictions of these models, one aims to maximise the amount of important features used to make successful predictions, based on which improved generalisation to unseen data which also contains these important features can be observed (X. Dong, Yu, Cao, Shi, & Ma, 2020). While approaches mentioned have been extensively studied, the use of annotator rationales requires more investigation to determine its effectiveness in increasing the OOD performance of a pre-trained transformers.

2.5 Increasing OOD performance with rationale-augmented models

The inductive bias instilled in rationale-augmented models is arguably more explicit than in models trained via other methods. As opposed to increasing variance in the data or the model to maximise predictive utility, annotator rationales can be used as an additional form of direct human supervision. In this way rationale

augmentation may enable the annotator to explicitly convey vital information that other approaches may be less sensitive to [4]. Although researchers have experimented with many rationale augmentation approaches, two of these methods have been most widely used, namely *multi-task learning* and *attention regularisation* (see Table 2). These methods are discussed in the following subsections.

2.5.1 Multi-task learning

The underlying idea of this approach is to train a model to perform two tasks simultaneously, such that the knowledge obtained from each task can be used to increase performance on the other (Y. Zhang & Yang, 2018). In the case of rationale augmentation, these two tasks are (1) predicting the label of a given document and (2) for each token in the sequence, predicting whether that token is an annotator rationale (Melamud et al., 2019; Pruthi et al., 2022; A. Sharma et al., 2020). Given the main goal of this approach generally is to correctly predict the class label of the document, the latter task can be seen as auxiliary to the former. To perform both tasks simultaneously, multi-task models make use of a modified loss function that combines the loss between the predicted and actual labels for both tasks. Melamud et al. (2019) found that the accuracy of a BERT model can be improved by 5% when the main task is supplemented by the auxiliary using a training set of 400 texts. Pruthi et al. (2022) observe less impressive, albeit noticeable results in this regard, achieving an increase in accuracy of 1.4% on a training set size of 1200. The results of both works were observed on the same dataset and task, namely sentiment analysis of IMDB movie reviews and accompanying annotator rationales collected by O. Zaidan et al. (2007). In contrast, M. Sharma, Zhuang, and Bilgic (2015) train a RoBERTa-based multi-task model to detect three types of empathy expressed in conversations taken from online mental health platforms. With a training set size of approximately 7500, this model was able to substantially outperform a 'vanilla' RoBERTa model on both accuracy and F1 metrics, with these improvements varying between 1 and 2.5% for accuracy and 1 and 4 points for F1-score, depending on which class and mental health platform was evaluated. While it is difficult to compare the results of these three works due to factors such as training set size, test set size and architectural choices, the benefits of multi-task learning are evident and potentially prevalent across different tasks. However, to my knowledge, these three studies are the only ones that have employed rationale augmentation via multi-task learning, all of which are based solely on ID test sets. Therefore, more research is needed to evaluate the effectiveness of this approach in creating robust models that can perform well across different domains.

2.5.2 Attention regularisation

Similar to multi-task learning, attention regularisation involves modifying the optimisation objective of the model (Kanchinadam et al., 2020; Pruthi et al., 2022; Barrett et al., 2018). However, in this case, the loss term that is added accounts for the difference between the attention weights assigned to a document's words and the corresponding annotator rationales. In this way, the aim is to encourage the attention mechanism of the model to give greater weight to words considered salient by the annotator. Pruthi et al. (2022) have demonstrated the effectiveness of this approach by increasing the accuracy of a BERT model trained on only 600 examples by 6% on a sentiment analysis task, achieved by supplementing each example with rationales. However, the increases in performance become less prominent as more training data is added, suggesting that this approach is most effective when labelled data is in short supply. Similar regularisation methods have been used to augment other neural architectures, such as feed-forward, convolutional and recurrent neural networks, also resulting in noticeable improvements over baselines that do not utilise rationales (Kanchinadam et al., 2020; Barrett et al., 2018).

Unlike the aforementioned papers in this section, Bao et al. (2018) take a different approach by assessing the performance of rationale-augmented models on OOD data. In their study, the researchers trained multiple models on different sentiment analysis tasks, using training sets comprised of beer reviews and test sets consisting of hotel reviews. Their findings indicate that the use of attention regularisation can lead to improved performance of LSTM architectures, contingent upon the OOD data used for evaluation. Specifically, the model demonstrated a 1.7% boost in accuracy on hotel reviews pertaining to cleanliness, while suffering a 0.97% decline on hotel reviews related to the quality of service. More recently, Stacey et al. (2022) have shown the gains in OOD performance from rationale augmentation to carry over to transformer-based architectures. They train their own variant of a rationale-augmented BERT model using attention regularisation on an NLI task.

[4] Testing this hypothesis would require implementing these other approaches as well as rationale augmentation and comparing each approach. However, doing so falls beyond the scope of this thesis.

This model is shown to make improvements to both ID (0.4%) and OOD (between 0.79 and 1.59%) performance.

Importantly, Stacey et al. (2022) perform a number of post-hoc analyses of the final [CLS] token of their BERT model. The initial analysis conducted by the authors demonstrates that their rationale-augmented model distributes its attention to tokens in both the premise and conclusion almost equally. In contrast, the baseline model (trained solely on the cross-entropy loss between the predicted and actual class labels) predominantly pays attention to the conclusion rather than the premise. This analysis highlights the potential of attention regularisation to help the model utilise information from both parts of the input, rather than relying too heavily on one aspect. In their second analysis, Stacey et al. (2022) show that their rationale-augmented model attends to tokens, such as "man", "woman" and "people", while their baseline model attends more to words, such as "a", "are" and "the". These results suggest rationale augmentation via attention regularisation also seems to assist the model in shifting its focus from irrelevant stop-words towards more meaningful words, such as nouns. Overall, these two analyses indicate that attention regularisation achieves the desired effect, namely encouraging the model to focus more on relevant features compared to spurious ones.

2.5.3 Attention regularisation vs Multitask learning

While comparing multi-task learning and attention regularisation on the same tasks is an interesting endeavour, the experiments conducted in this thesis have been limited to attention regularisation due to time constraints. While neither of the two is an obvious choice, attention regularisation edges out multi-task learning for a number of reasons. Firstly, attention regularisation has been the most widely employed technique among prior researchers (see Table 2), providing a broader literature base to draw upon in my own work. Secondly, the analyses carried out by Stacey et al. (2022) demonstrate that attention regularisation is a promising technique for training models to disregard spurious features. Lastly, as the only study I am aware of to compare the two approaches, Pruthi et al. (2022) showed attention regularisation to be more effective than multi-task learning, both requiring less training examples as well as yielding higher test performance, albeit on ID test sets. Therefore, attention regularisation was chosen for the experiments in this thesis, as it has a broader literature base to draw upon, has demonstrated promise in disregarding spurious features, and has been shown to be more effective than multi-task learning in the only known study comparing the two approaches.

3 Methods

The methods used to address the research questions, as defined in Section 1, are outlined in this section. All deep learning experiments were run on a single NVIDIA RTX 6000 GPU and all scripts were executed using Python 3.9.16[5]. Section 3.1 introduces the datasets and their relevant statistics. This section is followed by a discussion on the preprocessing steps in Section 3.3. The BERT model, chosen for this study due to its widespread use in previous work on rationale augmentation (see Table 2), is detailed in Section 3.2. Section 3.4 describes the process by which the rationales are incorporated during the fine-tuning of BERT. Benchmarks for this study are established by comparing the main rationale-augmented model against three baseline models, as presented in Section 3.4.3. Finally, Section 3.6 discusses how similarity between attention weights and the rationales is computed.

3.1 Data

This section provides details on the ID and OOD datasets used to answer SQ1, namely

> *Does attention regularisation affect the OOD performance of pre-trained transformers similarly for sentiment analysis and emotion detection?*

The purpose of the following subsections is to explain how these datasets were curated and to justify how systematic differences between them can produce a natural distribution shift that can be used to evaluate the OOD performance of the rationale-augmented models I have trained.

3.1.1 Task 1: Sentiment analysis

Sentiment analysis, broadly defined as "*the computational study of people's opinions, attitudes, and emotions toward an entity*" (Medhat, Hassan, & Korashy, 2014), is a prevalent task for developing and testing rationale-augmented models. Given its broad applications, spanning areas such as consumer attitudes, investment trends, and potential security threats (Karlgren, Sahlgren, Olsson, Espinoza, & Hamfors, 2012), ensuring the robustness of sentiment analysis models to unseen data is crucial. The accuracy and reliability of these models hold substantial weight in driving sound decisions across these various contexts. For this task, two datasets were initially selected: a set of IMDB movie reviews as ID data and a set of Yelp restaurant reviews as OOD data. However, all models trained on the IMDB dataset unexpectedly showed a higher performance on the Yelp dataset than on the IMDB test set in initial experiments. This result suggested that the Yelp dataset was 'easier' for the models compared to the IMDB dataset. To probe the models in a more challenging OOD scenario, an additional dataset of beer reviews was incorporated for evaluation.

Furthermore, rationales in these datasets are represented as binary vectors of length n, with each entry corresponding to a word in the review. If an annotator considers the i-th token in a sequence as salient, then the i-th entry of the rationale vector is marked as 1. For instance, in the review "This food is delicious", the rationale vector [0,0,0,1] indicates that the word "delicious" is salient, while "This", "food", and "is" are not. Representing the rationales in this way provides a simple means to convey to the model which tokens in the input sequence to focus on.

IMDB (ID) (O. Zaidan et al., 2007)[6]
This dataset was selected because it is often used to train and evaluate rationale-augmented models in previous literature (O. Zaidan et al., 2007; Kanchinadam et al., 2020; Pruthi et al., 2022; Wang et al., 2022; Melamud et al., 2019; Y. Zhang et al., 2016). Hence, I can directly compare my results to these previous works. This dataset was created by O. Zaidan et al. (2007) using an existing collection of 1000 positive and 1000 negative movie reviews from IMDB, assembled by Pang, Lee, and Vaithyanathan (2002). To ensure diverse authorship, a maximum of 20 reviews from a single author were included, resulting in a dataset with reviews from 312 different authors. This approach was designed to prevent classifiers from converging to sub-optimal solutions due to inherent biases in a particular writing style. However, O. Zaidan et al. (2007) noted a limitation in the dataset - Pang et al. (2002) included only reviews with high or low ratings[7]. Therefore, the IMDB dataset may

[5] All code and data can be found at https://git.science.uu.nl/8474752/Thesis-Michael

[6] The link to the IMDB dataset provided in O. Zaidan et al. (2007) is no longer functional. However, details of this dataset can be found at https://github.com/tensorflow/datasets/blob/master/docs/catalog/movie_rationales.md

[7] The movie reviews did not follow a specific rating convention. Therefore, Pang et al. (2002) only selected reviews that were accompanied by a numerical rating system, such as a five-star rating system.

not adequately represent the nuanced sentiments typically encountered in real-world scenarios.

O. Zaidan et al. (2007) tasked a single annotator with highlighting salient words in 1800 of the 2000 reviews that Pang et al. (2002) collected. The words were chosen based on the original sentiment labels provided by Pang et al. (2002). To gauge the degree of consensus among different annotators, O. Zaidan et al. (2007) conducted a preliminary analysis. In this analysis, 150 of the original 2000 reviews were assigned to four distinct annotators, asking them to underline the features they considered salient. The agreement level among individual annotators varied, with 40% to 80% of tokens highlighted by one annotator also selected by another. This variation in agreement suggests that individual preferences can significantly influence the selection of salient features. Nevertheless, the overall consensus—measured by the percentage of tokens marked by one annotator and at least one other—was considerably higher. This finding indicates a degree of consistency in the overall annotation process, and the highlighted words are likely to accurately reflect the sentiment of the reviews. However, it is worth noting that the annotators involved in this preliminary analysis are not the same as the one who provided rationales for the main dataset. As a result, the generalisability of the findings from this subset of 150 reviews to the entirety of the dataset remains an open question and should be taken into consideration when interpreting the results.

The size of the training, validation, and test sets directly impacts the ID and OOD performance of the models. Consequently, it is crucial to choose these splits carefully. Among the works related to rationale augmentation, only three provide usable information, as others either use larger datasets, train their models on different tasks, or do not specify the data splits used. Wang et al. (2022) split the data evenly into sets of 600. In contrast, Melamud et al. (2019) chose a larger training set size of 900, with validation and test sets of 450 each. While Pruthi et al. (2022) do not specify how they split this dataset, they report using training set sizes of up to 1200. Presumably, the corresponding validation and test sets would each contain 300 instances. For this work, I have opted for a 1200:300:300 split. This choice facilitates direct model comparison with Pruthi et al. (2022). Moreover, having a larger number of training examples provides me with more flexibility to experiment with various training set sizes, allowing a closer investigation into the impact of dataset size on OOD performance.

Yelp (OOD) (Sen et al., 2020)[8]
This dataset originally consisted of 5,000 restaurant reviews taken from Yelp.com[9], with ratings ranging from one to five stars. Like the IMDB dataset, reviews with one or two stars were classified as negative, while those with four or five stars were positive. Neutral, three-star reviews were omitted as they don't clearly represent either sentiment. Each review was evaluated by three different annotators from Amazon Mechanical Turk, meaning each appears three times in the dataset. Annotators classified these reviews as 'Positive', 'Negative', or 'I don't know', highlighting words and phrases to justify their classification. As a result, the dataset comprises a total of 13852 reviews, including duplicates[10].

To ensure careful highlighting, Sen et al. (2020) took several measures. Firstly, they noted that the number of highlighted words and the time taken to highlight increased with review length, suggesting that annotators make informed selections rather than randomly choosing words. Secondly, the authors visually confirm that the annotators select meaningful portions of the text. Finally, a pilot study is conducted in which eight annotators employ two different strategies for highlighting words: *read-first*, which involves reading the entire text first before highlighting words, and *free-style*, which gives annotators the freedom to highlight words as they read. The *read-first* approach is shown to yield higher inter-annotator agreement (unclear how they calculate this) regarding highlighted words as well as higher accuracy classifying the ground-truth label, making it the preferred approach for their main study.

Although the *read-first* strategy is effective at extracting high-quality rationales, disagreements between annotators are still prevalent in the Yelp dataset. For example, consider the review depicted in Figure 3.
This example underscores two important challenges associated with the rationales, as identified by Sen et al.

[8]The Yelp dataset can be found at https://davis.wpi.edu/dsrg/PROJECTS/YELPHAT/index.html

[9]The link to the original dataset provided by (Sen et al., 2020) (2020) (https://www.yelp.com/dataset/challenge is unfortunately no longer functional. Therefore, it is impossible to provide any details about these reviews that (Sen et al., 2020) (2020) themselves do not provide)

[10]Sen et al. (2020) claim a dataset of 15000 reviews. However, only 13852 are available on their GitHub page. Given that these reviews were divided into eight subsets, one of which is missing, it is assumed that the remaining 1148 reviews have been omitted.

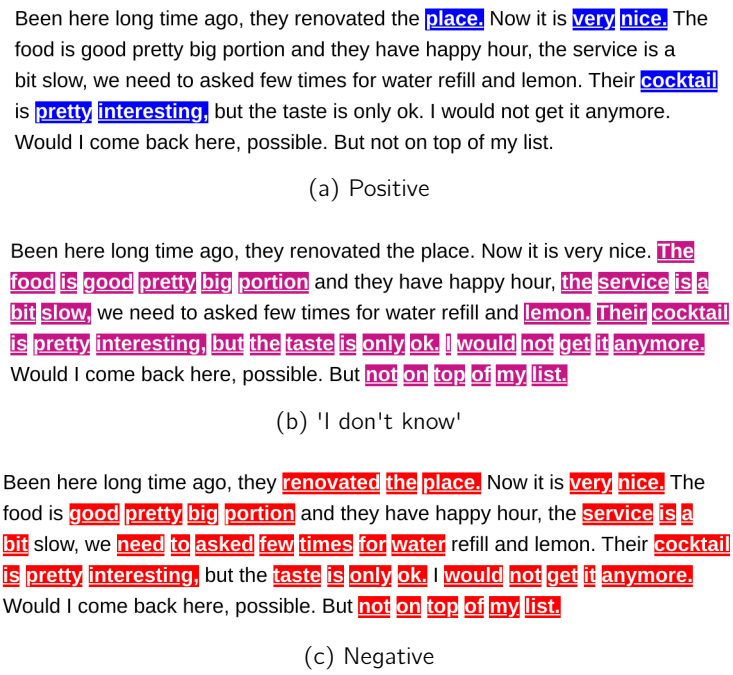


Figure 3: An example of a review in the Yelp dataset in which all three annotators provide a different label, namely 'Positive' (top), 'Negative' (bottom) and 'I don't know' (middle).

(2020) themselves. First, many reviews consist of a blend of positive, negative, and neutral sentiments. This can lead to disagreements among annotators on the overall sentiment of the document. As demonstrated in Figure 3, each of the three annotators selected a different label: 'Positive', 'Negative', and 'I don't know'. Second, despite adhering to the *read-first* annotation strategy, annotators may exhibit considerable variability in their highlighting methods. Some annotators adopt a more conservative approach, highlighting only specific words or phrases that align with their assessment, while others select all portions of the text they deem salient, irrespective of their sentiment categorisation. As shown in Figure 3, the annotator who labelled the review as 'Negative' highlights sections containing both positive (e.g., 'good pretty big portion') and negative (e.g., 'would not get it anymore') sentiments. In contrast, the annotator assigning a 'Positive' label uses a more conservative approach, highlighting only five words.

To address this variability between annotators, all rationales per document are aggregated into a single vector by taking the mean rationale value per word. For example, suppose the review "This food is good" has the following rationale vectors associated with it: [0,0,0,1], [0,1,0,1] and [1,1,1,1], the final rationale vector after aggregation becomes [0.333, 0.667, 0.333, 1]. To avoid aggregating conflicting information, only rationales corresponding to the majority label agreed upon by the annotators were aggregated. However, in some cases, this majority label contradicts the gold label. In such cases it is unclear whether the majority label, and hence the corresponding rationales, provide valid information. To avoid confusion, all reviews to which this conflict applies were removed from the dataset. As a result, a total of 2842 reviews were removed, resulting in 11010 reviews remaining in the dataset. Another 6 non-English entries were also removed. After aggregating the rationales of the remaining entries and removing all duplicated reviews, the final dataset consists of 4281 reviews[11].

BeerAdvocate (OOD) (Bao et al., 2018)[12]
This dataset is derived from a dataset of beer reviews cultivated by Lei, Barzilay, and Jaakkola (2016)[13]. Each review is assigned separate ratings between 0 and 5 stars for four different aspects of the beer being reviewed,

[11]As annotators may be removed from the dataset based on conflicts with the majority label, larger rationale values may be achieved for reviews with less annotators. While this approach by itself would entail that some reviews would unfairly influence the model's training process more than others, all rationales per entry are normalised to sum to one before attention regularisation is applied.

[12]The BeerAdvocate dataset can be found at https://github.com/YujiaBao/R2A

[13]Lei et al. (2016) themselves derived their dataset from McAuley, Leskovec, and Jurafsky (2012), who collected more than 1.5 million beer reviews. However, no details are provided on the specific method or process that was used to collect these reviews from their website or from which time period these reviews originate.

namely 'look', 'aroma', 'palate', and 'taste'. The original purpose of this dataset was to train a multiclass classifier to classify each review as one of these aspects. In this way, reviews are selected for which the rating of one particular aspect is the least correlated[14], thereby reducing confusion and facilitating clear distinctions among the classes. As ratings of the 'taste' aspect remain highly correlated, this aspect is not taken into consideration, resulting in three subsets for the aspects 'look', 'aroma', and 'palate'. Additionally, Lei et al. (2016) normalise these ratings to values between 0 and 1[15], as the original ratings were often fractional values. In their own study, Bao et al. (2018) assign all reviews with ratings ≤ 0.4 and less to the negative class and reviews with ratings ≥ 0.6 to the positive class. Next, they randomly select 100 positive and 100 negative reviews for each of the three aspects and instruct five human annotators to provide rationales. Unfortunately, details of this annotation process are not provided. Therefore, it is unclear what instructions were given to the annotators or whether all annotators provided rationales for each document or whether all documents were divided among the five annotators. In any case, one set of rationales is provided per document.

As using the entire dataset created by Bao et al. (2018) (more than 100,000 reviews) would result in excessive computational time required for OOD evaluation, only the test splits provided by Bao et al. (2018) for all three beer aspects of this dataset, containing 12030 reviews in total, was used as the OOD dataset. Of these 12030 reviews, 5 non-English reviews were removed, while 121 reviews were found to occur more than once in the dataset. For example, consider the following review:

> *i must admit i was suprised to see this new offering from coopers as i have never heard of it, it pours a pale golden see thru champagne colour with a nice frothy one finger head, sparkling carbonation and bubbly clingy lacing, it has a green, grassy hop aroma with a malty background, it has a thin mouthfeel even thinner for the style for my liking and the taste isnt as pleasant as it looks it has tastes of watery hops with a feint grainish aftertaste not as great as i was hoping for but definately better than the other macro swill out there*

This text is present in both the 'look' and 'palate' test splits. As this text is assigned to the positive class in the case of the former category and the negative class in the case of the latter, it is unclear which of the two labels to discard. To avoid confusion, these duplicates were removed. As a result, the final dataset consists of 12025. For comparing the model's attention weights and rationales for the same input sequences, the subset of 600 reviews with rationales was used. 1 review was removed from this subset where the number of rationales did not correspond to the number of tokens in the review, while 53 duplicated reviews were removed, resulting in 547 reviews.

Comparison between IMDB, Yelp and BeerAdvocate datasets

The distribution shift between these three datasets can be studied from both a qualitative and quantitative perspective. Qualitatively, the Yelp and BeerAdvocate datasets are similar to the IMDB dataset in that all three contain reviews of a particular product and are all are labeled based on a numerical rating provided by the reviewer. However, despite these initial similarities, the datasets diverge in two important ways: (1) the platform, as the IMDB platform may attract different types of users to Yelp.com and BeerAdvocate and (2) the subject of the reviews, with IMDB concerning movies, Yelp concerning restaurants and BeerAdvocate concerning beer[16]. Quantitatively, as depicted in Table 3, both BeerAdvocate datasets have proportionally more tokens not found in the IMDB dataset compared to the Yelp dataset, suggesting that the content of the beer reviews may be further removed from the content of the IMDB dataset. Finally, the IMDB reviews are on average substantially longer than both OOD datasets.

Overall, these shifts may not just be caused by syntactic - due to differences in length and vocabulary - but also semantic elements, as reviews for movies, restaurants, and beers would involve different subjects, themes, and sentiments. However, understanding distribution shifts in NLP is a complex task, and the factors outlined here are possible influences rather than definitive causes. As such many other factors, such as (dis)similarities in

[14] For each aspect, the authors train a linear regression model to predict the rating of that aspect given the other three aspects. Reviews with low prediction error are removed from the dataset.

[15] The precise normalization method used is not reported.

[16] Changes in societal trends, product features, and language use over time can also significantly impact the contents of reviews. Therefore, another likely cause of distribution shift is a difference in time periods in which the IMDB, Yelp and BeerAdvocate reviews were written. While O. Zaidan et al. (2007) report the IMDB reviews being written before 2002, Sen et al. (2020) and McAuley et al. (2012) do not report the time period in which the reviews they collected were written.

word frequency or embedding representations of documents, could also contribute to these shifts. Regardless, the potential differences between the datasets listed in this section provide a foundation for evaluating the OOD performance of machine learning models.

Statistic	IMDB	Yelp	Beer (Rationales)	Beer (No Rationales)
Total Examples	1800	4281	547	12025
# Negative : # Positive	1.00	0.91	1.06	1.00
Vocabulary Size	51145	25577	5753	33048
Fraction Stopwords	0.41	0.40	0.37	0.38
Fraction Punctuation marks	0.04	0.04	0.04	0.04
Mean Text Length	398.16	70.50	128.41	137.93
Fraction Sub-tokens	0.05	0.08	0.03	0.023
Fraction Tokens not in ID dataset	NA	0.01	0.01	0.01

Table 3: Dataset statistics for the sentiment analysis task. The statistics on the IMDB dataset are based on its truncated version.

3.1.2 Task 2: Emotion Detection

In this thesis, this task involves classifying a text as one of five different emotions, namely fear, disgust, anger, joy and sadness. Creating automated emotion detection systems can have many practical applications, such as providing feedback to urban planners improve the well-being of inhabitants of smart cities (Guthier, Alharthi, Abaalkhail, & El Saddik, 2014), or serve as a cost-effective means to collect large amounts of data to study mental health disorders (M. Park, Cha, & Cha, 2012). As previous works have not done so, this experiment provides initial insights into the effects of attention regularisation on the performance of multi-label classifiers in both ID and OOD settings. To the best of my understanding, the Hummingbird dataset (Hayati, Kang, & Ungar, 2021) holds a distinctive position as it features each text annotated with multiple class labels and their associated rationale vectors. Therefore, this dataset provides an ideal base for the training and evaluation of multi-label classifiers that employ attention regularisation. Furthermore, a dataset coined "GoEmotions" serves as a suitable OOD dataset, as it has previously been used to evaluate the performance of classifiers trained on the Hummingbird dataset (Hayati, Park, Rajagopal, Ungar, & Kang, 2023).

Hummingbird (ID) (Hayati et al., 2021)[17]
This dataset was created to explore which cues BERT models rely on in a text to capture the manner in which one's thoughts, dispositions or emotions are expressed in that text, and whether these cues differ to those captured by humans. For example, the sentences "I hate this restaurant!" and "I am not really a fan of this restaurant" both convey that the restaurant in question is of low quality. However, the first sentence is written in a rude or frustrated tone, while the second is more polite. Hence, while a sentiment analysis model would arguably identify similar cues in both sentences (e.g. phrases such as 'I hate' and 'not really a fan'), a model trained to detect the politeness of a text may focus on cues that differ more between the two sentences. To this end, Hayati et al. (2021) collect a number of texts labeled according to eight different categories, namely politeness, sentiment, offensiveness, anger, disgust, fear, joy and sadness. These texts originate from four different existing datasets:

- **StanfordTreebank** (Danescu-Niculescu-Mizil, Sudhof, Jurafsky, Leskovec, & Potts, 2013): A corpus of requests taken from the Wikipedia community of editors and the Stack Exchange question-answer community. These requests are labelled as polite and not polite.
- **Sentiment Treebank** (Socher et al., 2013) A dataset in which each text is a single sentence taken from a movie review. The original movie reviews were collected from Rotten Tomatoes by Pang and Lee (2005). Similar to the IMDB, Yelp and BeerAdvocate datasets, these reviews are labelled as having either a positive or negative sentiment.
- A dataset of tweets compiled by Davidson, Warmsley, Macy, and Weber (2017). These tweets were extracted through the Twitter API using keywords and phrases considered to be hate speech. Accordingly,

[17]The Hummingbird dataset can be found at https://github.com/sweetpeach/hummingbird

these tweets are categorised as (1) hate speech, (2) offensive but not hate speech and (3) neither hate speech nor offensive.

- **SemEval 2018** (Mohammad, Bravo-Marquez, Salameh, & Kiritchenko, 2018): tweets collected via the Twitter API based on keywords for four basic emotions, namely anger, fear, joy and sadness. These texts are then annotated for the presence or absence of 11 different emotions. As Hayati et al. (2021) are only interested in the emotions fear, disgust, anger, joy and sadness, they presumably filter the SemEval 2018 data for these five emotions. However, this filtering step is not explicitly stated.

Hayati et al. (2021) extracted the texts that were most distinguishable in terms of the tone expressed by training one BERT model for each of the four datasets to classify the texts based on their original labels. Each model was then applied to each of the validation sets of the four datasets (4 x 4 = 16 total evaluations). Next, a total of 500 texts assigned the highest confidence scores were selected. Specifically, these texts consisted of the 50 most polite texts, 50 most impolite texts, 50 positive texts, 50 negative texts, 100 offensive texts, and 200 emotional texts (40 for each emotion). However, as mentioned, these texts do not necessarily have to have been selected based on their initial class label. For example, a tweet from the Stanford Treebank corpus that was annotated as impolite may have been assigned a high confidence score by the BERT model trained to identify hate speech.

After these texts were selected, 622 annotators were hired from Prolific to label and provide rationales for each text-emotion pair (500 texts x 8 labels = 4000 pairs). Each unique pair was assigned to three different annotators, who were asked whether the emotion in question was represented in the corresponding text and to highlight the words they based their answer on. The final label provided in the dataset was the majority vote of these three annotators. Finally, Hayati et al. (2021) used the highlighted words to calculate a 'human perception score' H for each word as follows

$$H(w_i) = \frac{\sum_{j=1}^{\#annotators} h_j(w_i)}{\#annotators} \tag{1}$$

Where w_i is the i-th word of an input sequence and h is a function that indicates whether the j-th annotator considered that word to be a cue that either supports or contradicts an emotion being present in the text. h is defined as

$$h_j(w_i) = \begin{cases} 1 & \text{if the word is a positive cue for an emotion.} \\ -1 & \text{if the word is a negative cue for an emotion.} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

The final importance score of each word is thus the mean of the importance scores assigned by each annotator, resulting in a rationale vector that is similar to those calculated for the Yelp dataset[18].

Two important modifications were made to the Hummingbird dataset to make it suitable for multi-label classification. Firstly, to limit the scope of this experiment specifically to emotion detection, only the labels corresponding to the emotional categories, namely 'fear', 'disgust', 'anger', 'joy' and 'sadness', were retained. Afterwards, each text was assigned a multi-hot encoding vector with five entries, with each entry indicating the presence or absence of one of the five emotions. For example, if a text is labelled 'fear' and 'disgust', it is assigned the label [1,1,0,0,0] (i.e. 1 if the emotion is present and 0 otherwise).

Secondly, careful considerations are necessary to split the Hummingbird dataset into training, validation, and test sets due to its unique composition and label distribution. The original study by Hayati et al. (2021) did not split this dataset, instead using the test sets of the four original datasets that compose Hummingbird. However, those sets cannot be used in this experiment, as they lack labels for all five emotional categories. With no precedence for splitting this dataset and no method to include additional data for validation and testing, a

[18]The key difference between the Hummingbird and Yelp rationales lies in how they handle annotators whose views conflict with the majority label: such annotators are included in the Hummingbird dataset, but I have excluded them from the Yelp dataset. The concern lies in the calculation of feature importance, where the same highlighted features by two annotators could potentially receive an average importance score of zero due to different class labels, even though both annotators might consider those features to support the same class.

60:20:20 split ratio is adopted. However it is crucial to observe that the 500 texts of the dataset do not evenly distribute the five emotional class labels. For instance, 'disgust' appears much more frequently than 'fear' (refer to Table 5). Hence, a random split could inadvertently skew class distributions among the training, validation, and test sets. To maintain a consistent class distribution throughout each split, an iterative stratification algorithm based on the work of Sechidis, Tsoumakas, and Vlahavas (2011) and Szymański and Kajdanowicz (2017) was implemented. This implementation was accomplished using the `scikit-multilearn` Python library (Szymański & Kajdanowicz, 2017). By iteratively allocating the texts among the three splits until the class distribution of each split mirrors that of the original dataset, this algorithm ensures that all splits yield a more accurate representation of the model's performance across all classes is achieved, thereby reducing the potential bias towards over-represented classes.

GoEmotions (OOD) (Demszky et al., 2020)[19]
This dataset contains texts spanning 2005 to 2019 from a Reddit data dump. Subreddits with a minimum of 10,000 comments were selected, and deleted and non-English comments were excluded. To address potential biases, harmful subreddits were avoided, offensive content related to demographic groups was manually reviewed, and comments were filtered based on length. To ensure that only comments in which one or more emotions is expressed sufficiently, a pilot classifier trained on a subset of comments was used to analyse the sentiment and emotions expressed in these comments. Based on this analysis, comments that did not express a sufficient degree of emotion were removed. 58,000 of the remaining comments were randomly sampled for annotation. Three different annotators were instructed to assign one or more of 27 emotional categories[20]. Finally, all labels that are assigned by only a single annotator are removed, as well as texts that are not assigned any label by any annotator are also are removed, resulting in the final dataset consisting of 54,263 comments.

GoEmotions was further modified for this thesis. Specifically, 179 duplicated entries with the same labels as their original counterparts were removed. An additional 74 duplicates were found with different labels. These duplicates were removed from the dataset, but the corresponding labels were merged with those of the originals. For example, if a review occurs twice in the dataset and is labelled as 'fear' and 'sadness' in the two occurrences respectively, the final review was assigned both 'fear' and 'sadness'. Finally, 47104 texts were removed in which none of the five emotion categories of interest, namely 'fear', 'disgust', 'anger', 'joy' and 'sadness', were present. Entries where other labels were present alongside one or more of these five categories were not removed, but the undesired labels were deleted. These labels were then converted to a multi-hot encoding format in the same way as the Hummingbird dataset. The final subset of the GoEmotions dataset used for this thesis consists of 6964 texts.

Comparison between the Hummingbird and GoEmotions datasets
The Hummingbird and GoEmotions datasets differ noticeably from one another in many ways. Qualitatively, there are two differences of note. Firstly, the Hummingbird texts come from Wikipedia, Rotten Tomatoes and Twitter, while GoEmotions is sampled from Reddit. Similar to the sentiment analysis experiment, this difference in platforms between the two datasets may attract different types of users. Secondly, the quality of the class labels supplied by the annotators may differ between the two datasets. In the case of Hummingbird, the annotators label each text-emotion pair sequentially, while for GoEmotions the annotators were free to label a single text with one or more emotions from a pre-defined list of 27 emotions. Choosing from 27 labels, as opposed to just two, significantly amplifies the complexity of the annotation process, potentially leading to less accurate or incomplete labels compared to Hummingbird. Additionally, the annotators hired to compile the Hummingbird dataset have more opportunity to reflect on whether the correct choice was made, as they were required to justify their decision by highlighting relevant words and phrases.

Quantatively, both datasets differ notably with regards to both class labels and input features. Looking at Tables 4 and 5, Hummingbird proportionally contains more texts that have been assigned multiple labels compared to GoEmotions, meaning models trained on the former may be more prone to outputting multiple labels, which is unsuitable for a large proportion of GoEmotions. Furthermore, while both datasets contain class imbalances, this imbalance is more pronounced in Hummingbird. For example, 'disgust' is present in almost half of the Hummingbird texts, while constituting less than 15% of the GoEmotions dataset. Consequently, Hummingbird may bias models to predict dominant classes more often than is warranted when GoEmotions is used for eval-

[19]The GoEmotions dataset can be found at https://github.com/google-research/google-research/tree/master/goemotions
[20]It is not stated whether three annotators were recruited in total or that three unique annotators from a larger group were assigned to each document. The latter is assumed to be true given the large number of comments to be annotated

uation. Regarding input features, there are many word-pieces and full words not present in the ID dataset that are present in the OOD datasets for these datasets compared to those used for sentiment analysis. As a result, models trained on Hummingbird face a considerable number of words GoEmotions they have not encountered during training. Hence, models that rely on spurious correlations, such as keywords present in the training data, may be unable to perform adequately on GoEmotions.

Taken together, the two datasets differ substantially in many ways, spanning from their source platforms to the distribution of both labels and input features. These differences arguably play an important role in identifying models that are unable to learn general-purpose features in Hummingbird that can be applied to other, distinct datasets such as GoEmotions.

Statistic	Hummingbird	GoEmotions
Vocabulary Size	4050	15998
Fraction Stopwords	0.35	0.36
Fraction Punctuation marks	0.06	0.05
Mean Text Length	19.00	12.84
Fraction Sub-tokens	0.18	0.11
Fraction Single Labels	0.58	0.98
Fraction Word-pieces not in ID dataset	NA	0.04

Table 4: Dataset statistics for the emotion detection analysis task.

	Fear		Disgust		Anger		Joy		Sadness	
	H	G	H	G	H	G	H	G	H	G
% Labels	0.16	0.11	0.42	0.15	0.35	0.28	0.23	0.26	0.26	0.23

Table 5: Class-level statistics for the emotion detection analysis task. 'H' and 'G' stand for 'Hummingbird' and 'GoEmotions' respectively.

3.2 Model architecture

All models are trained on the well-known BERT architecture (Devlin et al., 2019) and implemented using Huggingface's `bert-base-uncased` pretrained model[21] from the `transformers` library (Wolf et al., 2020). BERT is pre-trained on a large corpus of novels (Y. Zhu et al., 2015) and Wikipedia articles. BERT employs two distinct pre-training objectives that enable it to encode complex linguistic information. Firstly, a percentage of the input tokens are masked or replaced with a random token. By learning to predict these tokens, BERT's self-attention mechanism leverages information on both the left and right of the masked token. Therefore, the parameters learned encode non-trivial bidirectional information about language (Devlin et al., 2019)[22]. Secondly, BERT learns associations between pairs of sentences by predicting whether a one sentence is likely to follow the other. This task helps BERT model various relationships and coherence between sentences in a corpus that can be useful for downstream tasks such as text classification and question answering. While other, newer attention-based architectures could also have been used for this thesis, the original BERT model is by far the most used for rationale augmentation (see Table 2). I have thus opted for this architecture to compare my results with prior research.

[21] https://huggingface.co/bert-base-uncased

[22] This is in contrast to other architectures such as GPT (Radford et al., 2018) that only leverage information to the left of the token to be predicted.

3.3 Preprocessing

3.3.1 Tokenisation

To convert the data into a format that can be processed by BERT, each input text is passed to Huggingface's `AutoTokenizer`[23] algorithm from the `transformers` library (Wolf et al., 2020). This algorithm segments complex words into smaller 'word-pieces', encodes these word-pieces and adds [CLS] and [SEP] tokens to the beginning and end of these sequences respectively. While there are other preprocessing methods that could have been applied, such as stemming and lemmatisation, these methods are not appropriate, as the resulting abbreviated tokens may no longer represent features that were originally considered salient by the annotator.

3.3.2 Assigning rationale vectors to tokenised sequences

All word-pieces were assigned the same rationale value as their parent word. More specifically, suppose some text T consists of the words $w_1, w_2, \ldots, w_n$ and each $w_i \in T$ is converted into the word-pieces $t_{i,1}, t_{i,2}, \ldots, t_{i,m}$ by the tokeniser. Then, each word-piece $t_{i,j}$ that is created is assigned the same rationale value as w_i. Furthermore, in line with Stacey et al. (2022), all [CLS] and [SEP] tokens are assigned a value of 1. To illustrate the process of converting word-level rationales to token-level rationales, suppose the sentence 'I like this restaurant' is assigned the rationale vector [1,1,0,0] and is converted to the word-pieces '[CLS]', 'I', 'like', 'this', 'rest', '##au' and '##rant', '[SEP]'. Then, 'rest', '##au' and '##rant' are assigned the same value as their parent word 'restaurant'. Hence, including the [CLS] and [SEP] tokens, the resulting rationale vector corresponding to the tokenised sequence is [1,1,1,0,0,0,0,1]. Finally, in line with Stacey et al. (2022), the token-level rationale values are normalised to sum to one using the following equation:

$$r'_i = \frac{r_i}{\sum_{j=1}^{N} r_j} \tag{3}$$

Where r_i is the rationale value of the i-th token in a given tokenised sequence and N is the length of the tokenised sequence. By applying these preprocessing steps, one ensures that the word-level information encoded in the original rationales can be used to supervise the model's attention weights at the token level during fine-tuning.

3.3.3 Adhering to constraints on input sequence length

A drawback of the BERT architecture is that it allows each input sequence to contain a maximum of 512 tokens. As the reviews in the IMDB dataset are frequently significantly longer than 512 tokens, a strategy must be applied to ensure that such reviews can be processed by BERT. To my knowledge, only two previous studies that apply some form of rationale augmentation to BERT models explicitly address this problem, namely Wang et al. (2022) and Melamud et al. (2019). Both studies convert each sentence into embedding representations using an average pooling layer, which is effective in reducing computational complexity during fine-tuning and adheres to the maximum input sequence length constraints of the BERT architecture. However, this pooling approach cannot be used for attention regularisation because the attention matrices computed in the transformer layers no longer correspond to specific tokens. As a result, it becomes impossible to calculate the loss between attention scores and token-level rationales. As no further works can be drawn from, the default truncation method adopted by Huggingface's AutoTokenizer for BERT was applied, namely selecting the first k word pieces[24]. To retain the maximum amount of information in the reviews, k is set to 512.

3.4 Fine-tuning

3.4.1 Loss functions and optimisation objectives

In general, the fine-tuning procedure in this study involves calculating the loss L_{labels} between the predicted and actual labels for a given forward pass and propagating this loss backwards throughout the entire model. Similar

[23] https://huggingface.co/transformers/model_doc/auto.htmlautotokenizer

[24] Several studies have tested different truncation strategies (e.g. Sun, Qiu, Xu, and Huang (2019), Sheng and Yuan (2021) and Mutasodirin and Prasojo (2021)). However, an exhaustive review to select an optimal strategy is not necessary for this thesis, as other optimisation approaches are secondary issues. Hence, though not necessarily optimal, selecting the first k tokens serves as a sufficient starting point.

to most other works, the binary cross-entropy function (Murphy, 2012) is used to calculate this loss. Therefore, L_{labels} is defined by the following equation.

$$L_{labels} = -(y \log(p) + (1 - y) \log(1 - p)) \tag{4}$$

An additional loss term L_{att} is computed for all models to which attention regularisation is applied. Following Stacey et al. (2022), this term is computed as the mean squared error between the attention scores and rationale values for each token in a given input sequence. This computation is performed as follows:

$$L_{att} = \frac{1}{H} \sum_{h=1}^{H} \sum_{i=1}^{N} (a_{h_i} - r'_i)^2 \tag{5}$$

Where H is the number of attention heads, N is the number of tokens in a given input sequence, a_{h_i} is the attention weight of the [CLS] token in attention head h corresponding to the i-th token in an input sequence, and $r'i$ is the value in the corresponding rationale vector, normalised using Equation 3. Furthermore a_{h_i} is defined by

$$a_{h_i} = \frac{\exp\left(q_{h_{\text{CLS},i}}^{\top} \cdot k_{h_i} / \sqrt{d_k}\right)}{\sum_{j=1}^{n} \exp\left(q_{h_{\text{CLS},j}}^{\top} \cdot k_{h_j} / \sqrt{d_k}\right)} \tag{6}$$

This formula is based on the self-attention mechanism introduced by Vaswani et al. (2017). The query (q) and key (k) vectors are both computed within the final attention layer for each attention head. The dimensionality of the key vector is denoted by d, and the number of attention heads is represented by n. In this context, attention scores are specifically computed for the [CLS] token with respect to other tokens, as [CLS] token serves as a representation of the entire input sequence (Devlin et al., 2019). Put together, the overall optimisation objective for rationale-augmented models is

$$\text{argmin}_W \, L_{labels} + \lambda L_{att} \tag{7}$$

Where λ is a tunable hyperparameter and W is the set of weights in each attention head and across all layers of the model. Similarly, the optimisation objective for all models not augmented with rationales is

$$\text{argmin}_W \, L_{labels} \tag{8}$$

3.4.2 Extra considerations for emotion detection

In the context of multi-label classifiers that produce multiple labels for a single input sequence, additional considerations are required when computing the loss for the emotion classification problem. Similar to previous works, individual binary cross-entropy losses are computed for each label, and the mean of these losses is propagated backwards. However, there is a considerable class imbalance observed in the Hummingbird dataset compared to the GoEmotions dataset, particularly with the 'disgust' label. To address this imbalance, weights are computed per class using `sklearn`'s (Buitinck et al., 2013) `compute_class_weight`[25] function. These weights are determined as follows:

$$w_i = \frac{N}{K \cdot n_i} \tag{9}$$

Here, w_i is the weight computed for the i-th class, N is the total number of samples in the training set, K is the total number of classes and n_i is the number of samples in class i. With this formula, larger class weights are computed for classes that are less represented in the training data. Using these weights, the loss for non-rationale-augmented models trained on a multi-label dataset can be computed as:

[25] https://scikit-learn.org/stable/modules/generated/sklearn.utils.class_weight.compute_class_weight.html

$$\text{argmin}_W \frac{1}{K} \sum_{i=1}^{K} w_i L_{\text{labels}}$$

(10)

Furthermore, while there is limited prior research on the effects of attention regularisation for multi-label classification, incorporating rationales logically suggests computing L_{att} separately for each class and adding it to the binary cross-entropy loss for that specific class. Hence, the loss computed for all rationale-augmented models is given by:

$$\text{argmin}_W \frac{1}{K} \sum_{i=1}^{K} w_i (L^i_{labels} + \lambda L^i_{att}) \tag{11}$$

By computing the weights and attention loss separately for each class, the proposed approach ensures an equitable representation of the information proved by rationales across all classes, thereby reducing the likelihood of models favoring one class over the others. Hence, models are more likely to make a genuine effort to distinguish and classify instances for all classes. Although this approach cannot draw from previous empirical evidence and may benefit from further refinement in future studies, it serves as a promising starting point for exploring attention regularisation in multi-label settings.

3.4.3 Models

The main model used in this thesis, which I call **CE + R** (Cross-Entropy + Rationales), was fine-tuned using Equation 7 and forms the basis of testing the attention regularisation approach. However, to gain a proper understanding of the effectiveness of this approach, CE + R was compared to three baselines, each of which was designed to rigorously evaluate the effect each component of attention regularisation has on BERT's OOD performance. Testing each component individually helps us ascertain whether any observed benefits can be attributed to a particular component or their collective influence. Attention regularisation consists of three main components: L_{labels}, L_{att}, and the specific rationale vectors $\{R_1, R_2, \ldots, R_N\}$ that are used to compute L_{att}. In pursuit of this objective, three baselines were designed in order to isolate and evaluate the contribution of each of these components. These models based on modifications of the `bert-base-uncased` model, implemented using Pytorch (Paszke et al., 2019). The baselines are listed below.

1. **BERT-base**. This first baseline is a standard BERT classifier fine-tuned using only L_{labels}. By comparing this baseline to a model trained on both L_{labels} and L_{att}, one can straightforwardly determine whether the addition of L_{att} yields any benefits to OOD performance. However, it is unclear whether any observed differences in performance between this baseline and CE + R can be attributed to the specific rationales used to compute L_{att} or to the interaction between L_{labels} and L_{att} to compute the final loss of CE + R. The following two baselines address this ambiguity.

2. **Only rationales**. To establish the impact that the rationales themselves have, this baseline was trained by masking out all words that were not highlighted by human annotators. Specifically, the binary rationale vector was used in place of the original attention mask designed to mask out all [PAD] tokens present in input sequences shorter than the longest input sequence[26]. However, this baseline is not applicable to the emotion detection experiment, as there are no words that are considered a rationale by all five rationale vectors for any given sentence. Consequently, if this baseline were to be applied, all words would need to be masked, resulting in a lack of training information. Hence, this baseline approach is not used for the emotion detection experiment.

3. **CE + F** (Cross-Entropy + Flipped Rationales). This baseline was created to disentangle the influence of L_{att} itself from the specific token-level human supervision, allowing for a more comprehensive assessment of the effectiveness and contribution of the attention regularisation approach. Specifically, this model is

[26] all [PAD] tokens remain masked.

trained in the same fashion as CE + R (i.e. Equation 7). However, different to CE + R, each $r_i \in R$ (Equation 3) is now set to 1 if the corresponding token $s_i \in S$ not *not* highlighted as salient by the annotator. In the case of emotion detection, extra considerations must be made, as the Hummingbird dataset contains rationale vectors with continuous values that indicate how many annotators agree on a word being salient. In these cases, if r_i is greater than zero, it is set to zero. While doing so removes the nuance of multiple annotators, there is no other clear method to flip the rationales that can preserve this nuance. By isolating L_{att}, it becomes possible to evaluate the extent to which the regularisation term itself contributes to the model's improved performance independent of the specific token-level human supervision used.

3.4.4 Hyperparameters

Due to the vast number of possible hyperparameter configurations that can and have been explored in previous works, it is computationally infeasible to conduct an exhaustive grid search across the entire hyperparameter space. Avoiding this exhaustive search is justified as the primary goal of this thesis is not to improve on the current state-of-the-art performances, but to answer the research question formulated in Section 1.5, namely

RQ: How does attention regularisation affect the OOD performance of pre-trained transformers?

Therefore, a number of concessions have been made to ensure computational feasibility while still providing meaningful results. In line with this reasoning, the majority of the hyperparameters (e.g. embedding layer size and activation function) were fixed to the default settings provided by `bert-base-uncased`. In addition, all models were trained with a batch size of 8[27] and the Adam optimiser (Kingma & Ba, 2014)[28].

Three different hyperparameters were tuned via a grid search: the regularisation term λ, the number of attention heads used to compute L_{att} as defined in Equation 5 and the learning rate[29]. λ is arguably the most important hyperparameter to tune for this thesis, as it directly controls the degree to which L_{att} influences parameter updates of the model during fine-tuning. In their work, Stacey et al. (2022) select λ from the range [0.2, 1.8], with increments of 0.2. For simplicity, I have opted to sample λ from the same range, but with increments of 0.4. This approach provides a more economical way of exploring the hyperparameter space while still providing a reasonable overview of the effects of L_{att} on the training procedure when λ takes on both smaller and larger values.

In addition to λ, the attention heads that are regularised also play a vital role in the effects of L_{att} on the overall loss. As described by Stacey et al. (2022), additional supervision through L_{att} can sometimes hinder specific attention heads' ability to capture relevant features, resulting in negative impacts on overall model performance. To this end, Stacey et al. (2022) first supervise each attention head individually and rank each head according to the the observed increases in performance. During fine-tuning, the top K heads are selected for supervision, where K is tuned from the values 1, 3, 6, 9 and 12. For simplicity, the attention heads are not ranked in this thesis. Instead, the first K heads are used to compute L_{att}, where K is tuned from the set $\{2, 6, 10, 12\}$. In this way, both smaller and larger proportions of the total attention heads are represented in the hyperparameter space. Although this approach may not provide the same level of detail as ranking and supervising individual attention heads, it still provides valuable insights into the potential effects of different attention heads on the fine-tuning procedure.

The two hyperparameters discussed above are only applicable to two of the four models that are to be trained, namely CE + R model and CE + F. Without optimising the hyperparameters of the other two baselines for which attention regularisation is not applicable, one could argue that comparisons between the four models unfairly favor the two for which hyperparameter configurations are evaluated. For a fairer comparison, the learning rate was also tuned. While Stacey et al. (2022) and Pruthi et al. (2022) are the main sources of inspiration for this

[27] I attempted to experiment with larger batch sizes, such as 16, 32 and 63. However, using batch sizes of larger than 8 resulted in GPU memory errors

[28] Although Stacey et al. (2022) and Pruthi et al. (2022) are the main works that were used to inform the experimental setup of this thesis, the specific optimiser used is not mentioned in either work. As the next closest work, the same optimiser used by Mathew et al. (2021), namely Adam, was employed.

[29] As the emotion classification experiment involves computing the loss for five classes separately, both λ and the number of attention heads could be tuned separately for each loss term. However, for simplicity and computational efficiency, these hyperparameters were tuned uniformly for each loss term.

thesis, they do not report the learning rates they used in their experiments. However, Mathew et al. (2021), who similarly apply attention regularisation to a BERT architecture, use a learning rate of 2e-5. Inspired by this choice, initial experiments explored learning rates between 0.005 and 1e-5, of which the values 1e-5, 2e-5 and 5e-5 yielded the highest validation accuracies across all baselines. Hence, for the main experiment, the learning rate is selected from these three values. While these initial experiments by no means guarantee the optimal learning rate to be found, they provide the BERT-base and the 'Only Rationales' baseline with a degree of flexibility, allowing for a fairer comparison with the other two models.

The primary focus of the hyperparameter search in this study is on the configurations controlling the influence of the term L_{att} on the total loss, as it is central to the attention regularisation technique being investigated. Given that neither λ nor the number of attention heads, which are crucial to the computation of L_{att}, are typically optimised in deep learning experiments, a tool with sufficient flexibility is required. For this purpose, the `sherpa` library (Hertel, Collado, Sadowski, & Baldi, 2018) was used, owing to its ability to customise the tuned hyperparameters. As such, in-depth exploration of the hyperparameter space regarding attention regularisation allowed for a profound understanding of its implications on model performance and behaviour. Even though tuning other hyperparameters might significantly affect the model's performance, this study specifically underscored those directly linked to the main research question, thereby maintaining focus on the essence of the attention regularisation approach.

3.4.5 Random seeds

Similar to hyperparameter selection, the choice of random seed can fundamentally influence the results of one's machine learning experiment. The order of training data greatly influences the model's performance, as it can lead to unwanted local minima. To mitigate this issue, data is randomly shuffled before each epoch to account for batch-wise gradient variations. However, the choice of random seed can substantially impact the computed gradients and subsequent parameter updates during training. To account for this effect, I followed the approach of previous works (Stacey et al., 2022; Bao et al., 2018) and experimented with six different random seeds, namely 4, 8, 16, 20, 24 and 32, when training all models. Huggingface's `set_seed` function[30] from the `transformers` library (Wolf et al., 2020) was used to ensure that these seeds were set across the random, numpy and torch libraries To avoid a considerable increase in computational cost, I first optimised the hyperparameters for a single seed (32) and then trained a single new model for each additional seed using those same optimal hyperparameter settings found. The overall performance of the model is then averaged over the six runs. While this approach may introduce some potential bias towards the seed that was used for hyperparameter optimisation, it helps to neutralise any confounding factors introduced by random variation into the experiment, allowing for a clearer assessment of the effects of attention regularisation on OOD performance.

3.5 Evaluating performance

The evaluation of a model's performance involves various considerations, starting with the selection of appropriate metrics. Since the sentiment analysis and emotion detection experiments involve binary and multi-label classification respectively, specific considerations are discussed in Sections 3.5.1 and 3.5.2 to determine the relevant metrics for each experiment.

3.5.1 Metrics for sentiment analysis

In binary classification settings, choosing performance metrics is relatively straightforward, focusing on the classifier's ability to distinguish between two classes. In this experiment, traditional metrics including accuracy, precision, recall, and F1 score were employed. While accuracy provides an overall correct classification rate, it may not offer a complete performance assessment. Hence, precision and recall were used to evaluate the models' ability to identify positive and negative instances. Precision measures true positive predictions relative to positive predictions, while recall measures true positive predictions relative to actual positive instances. Considering both precision and recall is important in real-life scenarios where correctly identifying instances of a specific class may outweigh overall accuracy. To obtain a comprehensive evaluation, the F1-score was used to assess the model's balance between precision and recall. Precision, recall and F1 scores were computed using the `sklearn` (Buitinck et al., 2013) library, while accuracy was computed using `evaluate` from the `transformers` (Wolf et al., 2020) library.

[30] https://huggingface.co/docs/transformers/internal/trainer_utils

3.5.2 Metrics for emotion detection

Evaluating performance in multi-label scenarios introduces complexities beyond those found in binary classification. These complexities arise because a multi-label classifier must effectively differentiate among multiple classes. As a result, the performance metric used for model selection during hyperparameter tuning must ensure a balanced contribution from each class to the overall performance of the model. As indicated in Table 5, the Hummingbird dataset exhibits class imbalances, with the label 'disgust' appearing much more frequently than the other four emotions. Hence, macro-averaging was employed for model selection to address these imbalances. Macro-averaging calculates performance metrics individually for each class and then combines these results. Furthermore, echoing the methodologies of previous studies that used BERT for multi-label text classification tasks (Adhikari, Ram, Tang, & Lin, 2019; Amin et al., 2019; Zahera, Elgendy, Jalota, Sherif, & Voorhees, 2019; Chen et al., 2022; X. Zhang, Song, Feng, & Gao, 2021), the F1-score was used to measure the performance of each class that was macro-averaged. Finally, as only reporting the macro-F1 score may not provide a thorough understanding of test set performance, the macro and micro averages of precision, recall, and F1-score, in addition to exact match accuracy, are all reported for both the ID and OOD test sets.

3.6 Comparing similarity between rationales and attention weights

Aside from performance metrics, a number of similarity metrics were selected to answer the second research question formulated in Section 1.5, namely

> *SQ2: To what extent does attention regularisation guide the attention mechanism of pre-trained transformers to align with salient features highlighted by human annotators?*

Three metrics were selected, namely AUC-ROC, AUC-PR and cosine similarity, to measure thise similarity and were implemented using `sklearn`'s `cosine_similarity`, `roc_curve`, `precision_recall_curve` and `auc`. The following subsections outline some important considerations that were made to select to ensure the appropriateness of these metrics.

3.6.1 Previous approaches to comparing human and machine attention

Selecting a suitable metric to compare human and machine-generated data hinges on the inherent characteristics of the data. Various researchers have adopted distinct methods based on this premise. For instance, Herrewijnen et al. (2021) utilized metrics such as the Jaccard index, precision, recall, and F1 to measure the similarity between binary human and machine-generated attention scores. On the other hand, Sen et al. (2020) opted for the area-under-the-curve (AUC-ROC) to compare binary human attention with continuous machine attention scores. Complementary research trajectories focus on comparing machine-generated data with continuous representations of human data, including FMRI and eye-tracking data. Methods such as regression analysis (Søgaard, 2016; Hollenstein, de la Torre, Langer, & Zhang, 2019), representational similarity analysis (Abdou, Kulmizev, Hill, Low, & Søgaard, 2019), and correlation metrics such as Pearson's and Spearman's r (Eberle, Brandl, Pilot, & Søgaard, 2022; Sood, Tannert, Frassinelli, Bulling, & Vu, 2020) have been instrumental in these comparisons.

3.6.2 Selecting the appropriate similarity metrics

The challenge in this thesis lies in selecting an appropriate metric due to varying rationale representations across datasets. The IMDB and BeerAdvocate datasets feature binary rationale representations, while the Hummingbird and Yelp datasets employ continuous ones. Whichever metric is selected necessitates a compromise. For example, applying Pearson's r would require transforming binary rationale vectors into a continuous representation. This transformation results in sparse, homogeneous vectors that may be unsuited to establish an effective linear correlation that is required for Pearson's r to produce meaningful results. Conversely, to use the Jaccard index, the attention scores would have to be binarised according to some threshold value. However, the binarisation of these attention scores is heavily dependent on the threshold used and may result in the loss of potentially vital information. Overall, AUC-ROC requires the least sacrifices, as only the continuous rationales of Yelp and Hummingbird datasets must be binarised, while the attention scores may remain continuous values. Furthermore, AUC-ROC has been already been effectively used in related studies (Sen et al., 2020; DeYoung et al., 2020; D. Zhang et al., 2021). For these reasons, AUC-ROC was used as the primary similarity metric for rationales and attention scores across all datasets.

Additionally, to strengthen the robustness of this experiment and possibly unearth new insights, two additional metrics were incorporated: AUC-PR and cosine similarity. These metrics, although not conventionally used in this context, possess unique strengths that can supplement AUC-ROC. AUC-PR, the area-under-the-precision-recall-curve, is especially beneficial for comparing highly skewed vectors. As seen in Tables 11 and 12, the rationale vectors can be highly skewed in favor of non-rationales, making AUC-PR a fitting choice. Cosine similarity, on the other hand, measures the cosine of the angle between two vectors. While AUC-ROC and AUC-PR measure the extent to which the attention scores can be used to differentiate between human-identified salient and non-salient features, cosine similarity can provide a unique perspective on how closely the rationales and attention scores align in their orientation in a high-dimensional space, regardless of their magnitude. Comparing the results of these three metrics adds a layer of robustness to this experiment. If these metrics, each with its own focus, show similar trends, we can be more confident in the validity of the insights each metric provides. Furthermore, the novelty of applying these metrics in this context may offer fresh perspectives, adding value to the existing body of research in this area.

3.6.3 Issues with computing similarity

It is important to note that vectors being compared via AUC-ROC, AUC-PR and cosine similarity are required to contain at least one non-zero element to produce meaningful results. Therefore, all texts that are associated with zero vectors must be omitted from this analysis. For IMDB, these zero vectors are predominantly found in the positive class, while being more evenly distributed throughout both classes for Yelp and BeerAdvocate (see Table 6). The presence of zero vectors is more concerning for Hummingbird. Specifically, 57, 37, 38, 58 and 53 zero vectors were found for each of the five emotional categories respectively, meaning a considerable proportion of the 100 test examples must be omitted from this analysis on the Hummingbird dataset. Despite this limitation, the chosen metrics remain valuable tools in providing initial insights into how attention regularisation can be used to align the BERT's attention mechanism with salient features identified by humans.

% Zero-Vectors	IMDB	Yelp	BeerAdvocate
Negative	0.33	0.26	5.48
Positive	5.00	0.14	4.75
Total	5.33	0.40	10.24

Table 6: Percentage of zero rationale vectors found for the sentiment analysis datasets.

3.7 Varying training data

I draw from Bao et al. (2018) and Pruthi et al. (2022) to answer the final research question, namely

> *SQ3: Is attention regularisation an effective method for reducing the amount of training data required to achieve a desirable level of OOD performance in pre-trained transformers?*

In the domain of training rationale-augmented models in low-resource environments, a disparity exists between different studies. For instance, Bao et al. (2018) use training sets containing between 20 and 200 examples, while Pruthi et al. (2022) opt for much larger training set sizes, namely 600, 900, and 1200. For a comprehensive comparison between these works and my own, and to explore the impact of attention regularisation across diverse training set sizes, varied training set sizes were selected based on both works. For the IMDB dataset, these sizes include 25, 50, 100, 200, 400, 600, 900, 1200, while for the Hummingbird dataset the sizes 25, 50, 100, 200, 300 were used. Furthermore, to ensure consistency in the evaluation of model performance across different training set sizes, all models were trained and assessed using the same procedure. This procedure, as detailed in Section 3.4, includes fine-tuning and hyperparameter optimisation across six different seeds, as detailed in Section 3.4.5. Moreover, each subset from the full training sets was selected randomly, with the specific randomness determined by the respective seed. Using unique random seeds to select distinct subsets from the full training set increases the likelihood of avoiding bias in the selected subsets that may provide disproportionately large or small representations of the resulting models' OOD performance. This approach is effective at avoiding this bias because the randomness introduced by the seeds helps prevent any inadvertent

systematic selection of data points, which could potentially lead to an over- or under-representation of certain patterns or features in the subsets used for training. Overall, these considerations enable the learning curves resulting from this approach to provide more accurate insights into the extent to which attention regularisation can be used to reduce the amount of training data required to achieve desirable OOD performance.

4 Results

This section provides details of the results of the sentiment analysis and emotion detection experiments outlined in Section 3. Aside from discussing the models' classification performance and the similarity between their attention scores and the rationales, a number of qualitative analyses of individual instances in the OOD datasets were also performed to provide more nuanced insights into the potential effects of attention regularisation on BERT's OOD performance. However, is it important to note that these analyses are not necessarily representative of these effects on the entire OOD datasets.

Furthermore, the abbreviations of each model used in these experiments are re-iterated below to avoid confusion:

- **CE + R** (Cross-Entropy + Rationales). This model is fine-tuned on the combined loss of L_{labels} and L_{att}, namely the cross-entropy loss of the predicted and actual labels as well as the loss between the model's attention scores for individual tokens in the input sequence and the rationales corresponding to those tokens.
- **BERT-base**. A 'vanilla' BERT classifier that uses only L_{labels} during fine-tuning.
- **Only Rationales**. This model is also only fine-tuned with L_{labels}, but all words in the input sequence that are not rationales are masked.
- **CE + F** (Cross-Entropy + Flipped Rationales. This model is also fine-tuned with both L_{labels} and L_{att}. However, in this case, the rationale vectors used to compute L_{att} are 'flipped', meaning all zeros are converted to ones and all ones are converted to zeros.

4.1 Results sentiment analysis

4.1.1 Performance when trained on the full training set

Table 7 provides a summary of the models' ID (Table 7(a)) and OOD (Tables 7(b) and 7(c)) performance when trained on the full IMDB training set of 1200 reviews. In line with the pilot studies, all models achieve much higher performance on the Yelp dataset than on the IMDB dataset. Both CE + R and CE + F outperform BERT-base across all metrics, indicating that the efficacy of attention regularisation extends to OOD data. Furthermore, 'Only Rationales' achieves the highest precision across all datasets, but much lower recall. As such, the model appears to assign the majority of texts to the negative class, but is highly proficient at correctly classifying the texts that it assigns to the positive class. The models' performances on the BeerAdvocate dataset are harder to interpret. Most importantly, CE + R outperforms both BERT-base and CE + F in terms of accuracy, but the latter models outperform the former in terms of F1-score. Hence, CE + R is more proficient at classifying instances overall, while BERT-base is more effective at classifying positive instances. 'Only Rationales' achieves the highest accuracy, but lowest recall and F1-score, again suggesting it to be biased towards assigning texts to the negative class.

Overall, these results indicate that the components of attention regularisation as described in Section 3.4.3, namely the cross-entropy loss between predicted and actual labels L_{labels}, the token-level loss between the model's attention scores and the rationales L_{att}, and the human identified salient features encoded as rationale vectors $\{R_1, R_2, \ldots, R_N\}$, each contribute to the improvements in OOD performance seen in CE + R compared to BERT-base. CE + F and CE + R both show signs of improvements over BERT-base on OOD data, suggesting that the addition of L_{att} to L_{labels} is beneficial in OOD scenarios. Furthermore, CE + R outperforms CE + F on both OOD datasets, which indicates that the human-identified salient features encoded in the the rationales $\{R_1, R_2, \ldots, R_N\}$ may more beneficial than their flipped counterparts when computing L_{att}. However, while CE + R clearly performs the best on Yelp, the results on BeerAdvocate are more mixed. Hence, strong claims about the effects of attention regularisation on OOD performance in general cannot be made.

4.1.2 Learning Curves

The learning curves depicted in Figure 4 reveal comparable trends among all four models on the IMDB, Yelp, and BeerAdvocate test sets. Echoing Pruthi et al. (2022), the advantages of attention regularisation become more apparent when models are trained on smaller data sets. Both CE + R and CE + F substantially outperform

(a) IMDB

Metrics	CE + R	BERT-base	Only Rationales	CE + F
Accuracy	0.846 ± 0.024	0.828 ± 0.032	0.660 ± 0.067	**0.860** ± 0.001
Precision	0.870 ± 0.032	0.846 ± 0.061	**0.950** ± 0.038	0.857 ± 0.039
Recall	0.802 ± 0.094	0.800 ± 0.104	0.308 ± 0.158	**0.853** ± 0.0422
F1	0.830 ± 0.038	0.813 ± 0.044	0.438 ± 0.188	**0.853** ± 0.006

(b) Yelp

Metrics	CE + R	BERT-base	Only Rationales	CE + F
Accuracy	**0.960** ± 0.006	0.950 ± 0.008	0.926 ± 0.033	0.951 ± 0.0064
Precision	0.953 ± 0.018	0.936 ± 0.025	**0.971** ± 0.016	0.929 ± 0.0163
Recall	0.972 ± 0.019	0.970 ± 0.016	0.886 ± 0.077	**0.982** ± 0.008
F1	**0.962** ± 0.006	0.953 ± 0.0072	0.924 ± 0.039	0.955 ± 0.0054

(c) BeerAdvocate

Metrics	CE + R	BERT-base	Only Rationales	CE + F
Accuracy	0.656 ± 0.016	0.650 ± 0.019	**0.660** ± 0.049	0.638 ± 0.009
Precision	0.632 ± 0.028	0.622 ± 0.026	**0.770** ± 0.049	0.602 ± 0.012
Recall	0.761 ± 0.045	0.779 ± 0.044	0.480 ± 0.174	**0.825** ± 0.025
F1	0.689 ± 0.005	0.690 ± 0.003	0.563 ± 0.152	**0.696** ± 0.004

Table 7: Mean performance of all models on the IMDB, Yelp and BeerAdvocate datasets trained across six different random seeds. Standard deviations are indicated with '±' and the highest performances per metric are boldfaced.

BERT-base in terms of accuracy when trained on 200 or fewer examples. As the training set is increased in size, the effectiveness of attention regularisation diminishes, with BERT-base closing the performance gap when trained on 400 texts or more. As the training set increases in size, all models—except for 'Only Rationales' on the IMDB test set—converge towards a similar mean accuracy. This convergence suggests that, given sufficient training data, each of these three models is capable of achieving comparable accuracy scores in OOD scenarios. Interestingly, the accuracy of Only Rationales' on both OOD datasets fluctuates drastically between random guessing (0.5) and satisfactory accuracy scores, before stabilising from 600 training examples onwards. These fluctuations imply that this model performs less effectively in low-resource environments. In summary, attention regularisation is a beneficial approach to reducing the quantity of training data needed for BERT to achieve desirable OOD performance, with the most substantial benefits observed for training sets of size 200 or fewer.

4.1.3 Similarity between attention weights and rationales

Models to which attention regularisation is applied also achieve higher similarity between attention scores and rationales. It is clear from the the similarity scores shown in Table 8 indicate that BERT-base achieves lower similarity scores for all three metrics compared to the other three models. Specifically, CE + R clearly outperforms the other three models for both AUC-ROC and AUC-PR metrics on the IMDB and Yelp datasets, while CE + F achieves the highest cosine similarity. However, the results on the BeerAdvocate dataset again tell a different story, with 'Only Rationales' achieving the highest similarity for all three metrics. Similar to the models' classification performance discussed in Section 4.1.1, the similarity scores on the BeerAdvocate dataset are much closer together, suggesting that the effects of attention regularisation observed on the IMDB and Yelp do not apply to BeerAdvocate.

Interestingly, the trend for cosine similarity diverges from the other two metrics, with 'Only Rationales' registering the highest cosine similarity across all datasets. However, the disparity between it and both AUC-ROC and AUC-PR, coupled with an apparent preference for the least performing model, calls into question the ro-

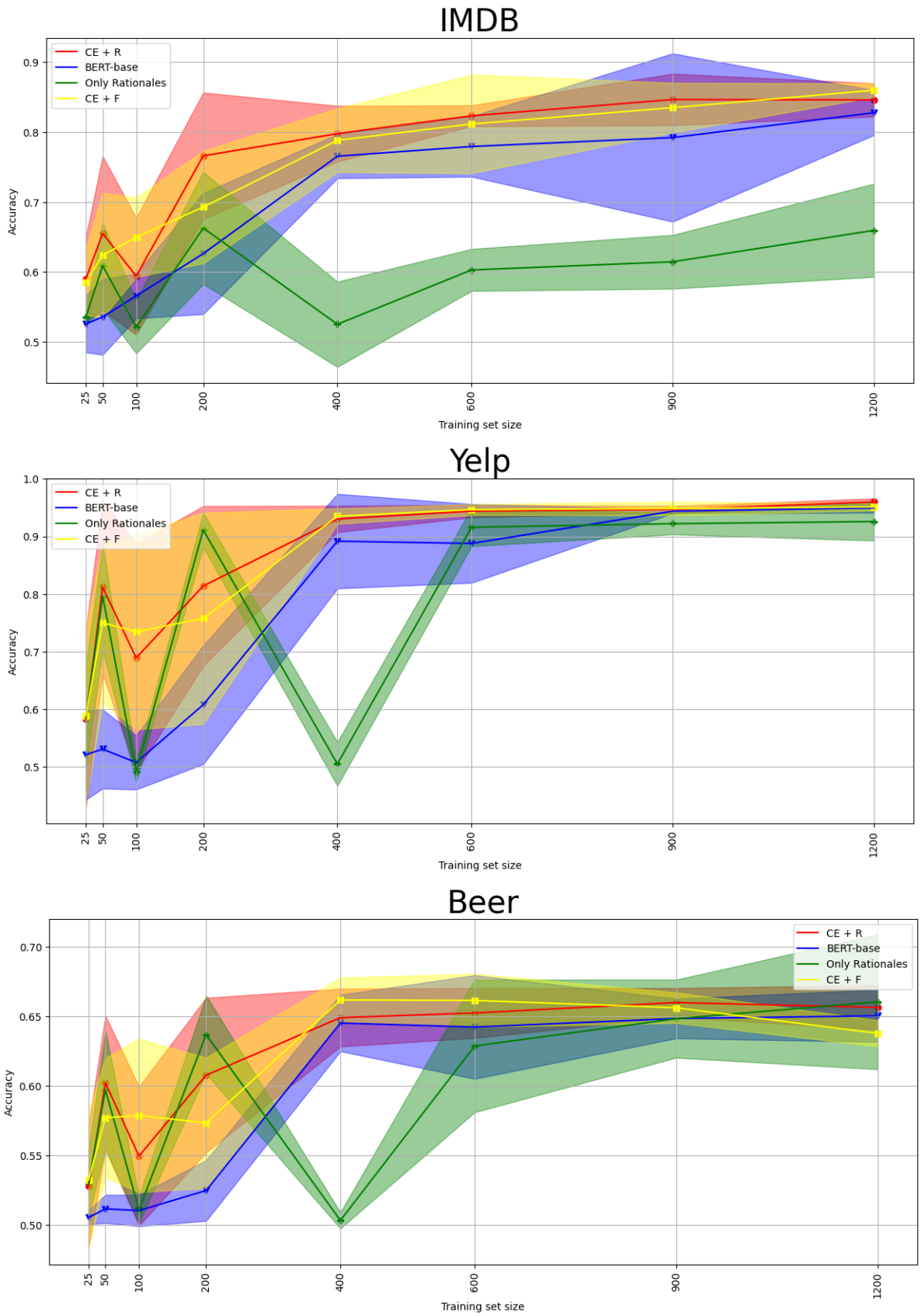


Figure 4: Learning curves for all models evaluated on the IMDB (top), Yelp (middle) and BeerAdvocate (bottom) datasets across six random seeds. Mean and standard deviations are plotted against the size of the training set used to train each model

bustness of cosine similarity as a metric for measuring the overlap between attention weights and rationales. The issue could be influenced by the impact of zero elements in the rationale vectors, which might distort the metric's values. Differences in the proportions of rationales to non-rationales and the corresponding variation in the number of zero elements can potentially conceal the actual overlap between attention scores and the genuine rationales. Consequently, using cosine similarity risks obscuring an accurate assessment of the model's focus on human-annotated words. Despite the anomalous results in cosine similarity, the CE + R model demonstrates significantly higher similarity with rationales compared to BERT-base. This finding suggests that attention regularisation can effectively enhance the alignment between attention weights and rationales.

Metric	CE + R	BERT-base	Only Rationales	CE + F
Cosine	0.097 ± 0.013	0.040 ± 0.008	**0.114** ± 0.078	0.084 ± 0.036
AUC-ROC	**0.772** ± 0.013	0.66 ± 0.040	0.549 ± 0.047	0.351 ± 0.018
AUC-PR	**0.191** ± 0.013	0.124 ± 0.015	0.103 ± 0.014	0.064 ± 0.001

(a) IMDB

Metric	CE + R	BERT-base	Only Rationales	CE + F
Cosine	0.232 ± 0.012	0.168 ± 0.007	**0.284** ± 0.080	0.229 ± 0.040
AUC-ROC	**0.668** ± 0.013	0.574 ± 0.021	0.578 ± 0.027	0.515 ± 0.012
AUC-PR	**0.336** ± 0.010	0.275 ± 0.01	0.281 ± 0.013	0.251 ± 0.005

(b) Yelp

Metric	CE + R	BERT-base	Only Rationales	CE + F
Cosine	0.080 ± 0.005	0.059 ± 0.0038	**0.146** ± 0.082	0.110 ± 0.032
AUC-ROC	0.470 ± 0.005	0.442 ± 0.009	**0.490** ± 0.032	0.443 ± 0.005
AUC-PR	0.151 ± 0.002	0.142 ± 0.002	**0.161** ± 0.023	0.140 ± 0.001

(c) BeerAdvocate

Table 8: Mean similarity scores between rationales and attention scores across six random seeds for all models on the IMDB (top), Yelp (middle) and BeerAdvocate (bottom) datasets. The highest similarity scores per metric are boldfaced and standard deviations are indicated with '±'. CE + R refers to the model that uses rationales to regularise its attention mechanism. BERT-base is a standard BERT model that uses only cross-entropy loss to update its parameters. 'Only Rationales' also uses only cross-entropy, but all words in each review that are not rationales are masked. Finally, CE + F is trained similarly to CE + R, but uses all words that are not rationales to regularise its attention mechanism

4.1.4 Qualitative Analyses

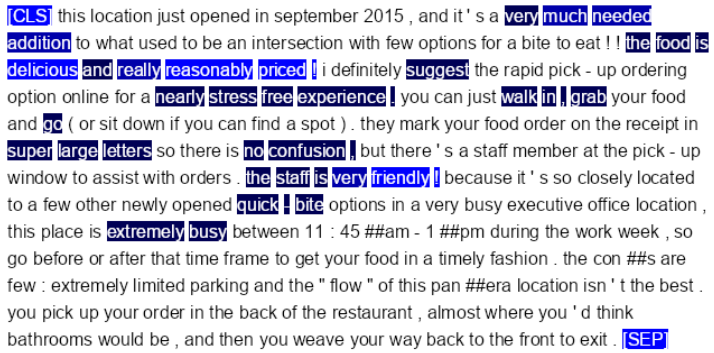
While CE + R surpasses BERT-base in performance across all three datasets, it is crucial to understand that the benefits of attention regularisation on OOD performance are not universal to all texts. To dig deeper, the analysis presented here considers two particular examples: (1) a review where CE + R accurately predicts the label, but BERT-base does not, and (2) a review where BERT-base gets the label right, but CE + R does not. Visual representations of the attention scores for each token in these reviews are offered in Figures 5 and 7. It is worth noting that these analyses are restricted to the attention scores of the [CLS] token in the model's final layer, thus providing a narrow view of the impact of attention regularisation on both similarity and performance.

The Yelp review depicted in Figure 5 is labeled as positive, but exhibits a certain level of ambiguity. While the reviewer praises the taste of the food ("*The food is delicious*"), they also mention the restaurant's "*extremely limited parking*". This mixture of positive and negative elements makes the overall sentiment of the review unclear. While both CE + R and BERT-base models consider words and phrases with both positive and negative sentiments, BERT-base mistakenly classifies the review as negative with relatively high confidence (0.8632), whereas CE + R accurately classifies the review as positive with relatively low confidence (0.6936).

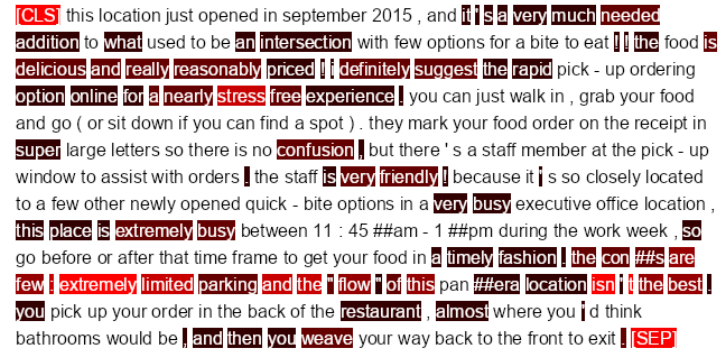
The primary distinction between the two models lies in their focus on specific words and phrases. CE + R places more emphasis on positive sentiment indicators, such as "*very much needed addition*" and "*for a nearly stressful experience*" compared to BERT-base. These positive phrases may contribute CE + R's decision to assign the text to the positive class.

Additionally, the positive cues used by CE + R, but not by BERT-base also make up a notable portion of the words highlighted by human annotators, shown in Figure 5 (a). Accordingly, CE + R achieves similarity scores of 0.1010, 0.9181 and 0.1763 for cosine similarity, AUC-ROC and AUC-PR respectively, while these scores are much lower for BERT-base, namely 0.0129, 0.3473 and 0.0233. These higher similarity scores potentially translate to an increased focus on salient features highlighted by human annotators in the case of the Yelp dataset. However, it is important to note that CE + R also considers the same negative portions of the text as BERT-base, indicating its recognition of salient features contributing to both positive and negative classes, rather than solely relying on the features highlighted by the annotator.

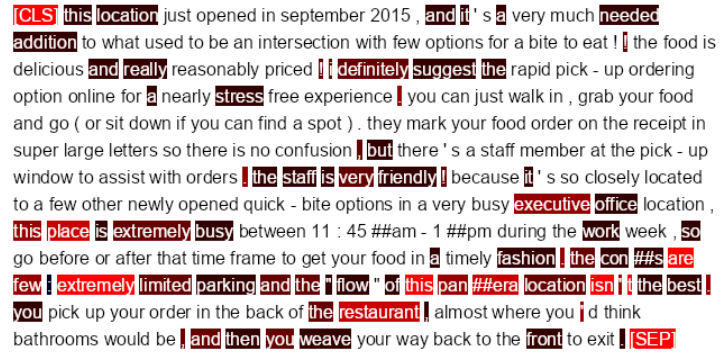
It would not be prudent to assert a relationship between similarity and performance based on a solitary instance; however, a Mann Whitney U-test (Mann & Whitney, 1947) between the AUC-ROC scores for the correct predictions and incorrect predictions implies the existence of such a relationship in the context of the Yelp dataset. As depicted in Figure 6, both CE + R and BERT-base tend to secure higher similarity scores on average when they make correct predictions, although this trend is less pronounced for BERT-base. However, additional investigations and thorough analyses are required to confirm this observation and to probe its applicability across other datasets, models and random seeds.



(a) Rationales



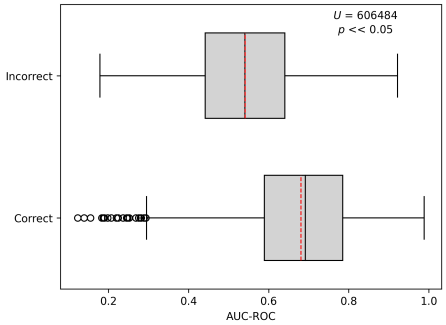
(b) CE + R. Predicts the correct label (1) with 0.6936 confidence



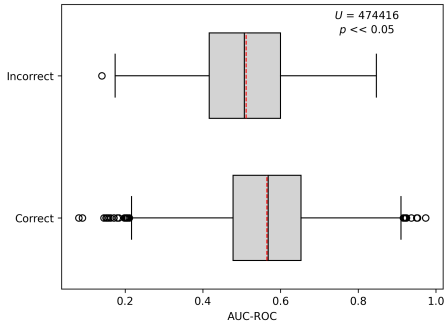
(c) BERT-base: predicts the incorrect label (0) with 0.8632 confidence

Figure 5: An example from the Yelp dataset were CE + R predicts the correct label and BERT-base does not. Tokens are highlighted according to the attention scores from the [CLS] token in the final layer. Brighter colors indicate higher attention scores

While the previous example suggests that higher similarity scores could potentially indicate better performance, this reasoning does not hold true in all cases, as exemplified by a BeerAdvcate review depicted in Figure 7. This review featured is labeled as negative, which is correctly predicted by BERT-base with a confidence score of 0.8186, but inaccurately predicted by CE + R with a confidence score of 0.9074. However, the attention scores of BERT-base do not provide clear indications of its focus on words highlighted by the annotator that CE + R overlooks, which would have led to the correct prediction. Furthermore, both models demonstrate comparable similarity scores, with CE + R scoring 0.2966, 0.4934, and 0.2190 for cosine similarity, AUC-ROC, and AUC-PR, respectively, and BERT-base scoring 0.2779, 0.4920, and 0.2243. Consequently, the disparity in performance between the two models cannot be attributed to the similarity score in this particular case. Overall, this example highlights two points: (1) attention regularisation may impede a model's performance in certain



(a) CE + R



(c) BERT-base

Figure 6: A visualisation of the similarity scores associated with correct and incorrect predictions of CE + R and BERT-base trained on a single random seed on the Yelp dataset. The means and standard deviations of the similarity scores for incorrect and correct predictions are plotted as box plots. Furthermore, 'U' indicates the U-statistic of the Mann-Whitney U-test. This test is a non-parametric approach to determine whether the means of two independent samples are significantly different from one another. In this analysis, these groups are AUC-ROC scores for correct predictions versus the AUC-ROC scores for incorrect predictions of CE + R and BERT-base.

scenarios, and (2) similarity and attention scores do not always serve as reliable predictors of a model's behavior and performance.

[CLS] we decided this one tasted a bit like ' gui ##ness light ' , poured a thing brown color with an impressive fr ##oth ##y head . nose was spot on , all dark and bold , with notes of dark chocolate . flavor wasn ##t bad , but the mouth ##fe ##el left a lot to be desired . watery through and through with a very forget ##table slightly bitter after ##tas ##te . recognizing that this is a fairly new operation , i will give this one a little slack , but just doesn ##t measure up to what iv ##e come to expect from a micro ##bre ##wed stout . the good news is that it was very session ##able , as it doesn ##t sit in the gut . so so [SEP]

(a) Rationales

[CLS] we decided this one tasted a bit like ' gui ##ness light ' , poured a thing brown color with an impressive fr ##oth ##y head . nose was spot on , all dark and bold , with notes of dark chocolate . flavor wasn ##t bad , but the mouth ##fe ##el left a lot to be desired . watery through and through with a very forget ##table slightly bitter after ##tas ##te . recognizing that this is a fairly new operation , i will give this one a little slack , but just doesn ##t measure up to what iv ##e come to expect from a micro ##bre ##wed stout . the good news is that it was very session ##able , as it doesn ##t sit in the gut . so so [SEP]

(b) CE + R: predicts the incorrect label (1) with 0.9074 confidence

[CLS] we decided this one tasted a bit like ' gui ##ness light ' , poured a thing brown color with an impressive fr ##oth ##y head . nose was spot on , all dark and bold , with notes of dark chocolate . flavor wasn ##t bad , but the mouth ##fe ##el left a lot to be desired . watery through and through with a very forget ##table slightly bitter after ##tas ##te . recognizing that this is a fairly new operation , i will give this one a little slack , but just doesn ##t measure up to what iv ##e come to expect from a micro ##bre ##wed stout . the good news is that it was very session ##able , as it doesn ##t sit in the gut . so so [SEP]

(c) BERT-base: predicts the correct label (0) with 0.8186 confidence

Figure 7: An example from the BeerAdvocate dataset were BERT-base predicts the correct label and CE + R does not.

4.2 Results Emotion detection

4.2.1 Performance when trained on the full training set

The findings of this experiment indicate that attention regularisation is also beneficial for emotion detection. Table 10 indicates that CE + R achieves the highest Macro F1 on the Hummingbird dataset, while also achieving both the highest macro and micro F1 on the GoEmotions dataset. Hence, CE + R is the most effective at balancing precision and recall when evaluated on GoEmotions. Additionally, CE + R achieves the highest macro and micro precision on both datasets, while BERT-base achieves the highest macro and micro recall. Finally, contrary to the sentiment analysis experiment, CE + F is overall less effective on both emotion detection datasets.

Table 9: Emotion detection results on the Hummingbird datasets

Metrics	CE + R	BERT-base	CE + F
Macro Precision	**0.609** ± 0.106	0.536 ± 0.036	0.481 ± 0.223
Micro Precision	**0.633** ± 0.072	0.566 ± 0.039	0.517 ± 0.246
Macro Recall	0.432 ± 0.046	**0.468** ± 0.029	0.326 ± 0.184
Micro Recall	0.473 ± 0.076	**0.532** ± 0.033	0.351 ± 0.205
Macro F1	0.454 ± 0.049	**0.462** ± 0.033	0.347 ± 0.173
Micro F1	0.532 ± 0.029	**0.547** ± 0.014	0.397 ± 0.197
Exact Match	**0.303** ± 0.077	0.267 ± 0.041	0.273 ± 0.042

Table 10: Emotion detection results on the GoEmotions datasets

Metrics	CE + R	BERT-base	CE + F
Macro Precision	0.451 ± 0.024	0.451 ± 0.049	**0.472** ± 0.021
Micro Precision	0.406 ± 0.069	0.373 ± 0.043	**0.478** ± 0.097
Macro Recall	0.496 ± 0.053	**0.520** ± 0.047	0.401 ± 0.144
Micro Recall	0.528 ± 0.058	**0.554** ± 0.056	0.425 ± 0.142
Macro F1	**0.405** ± 0.071	0.398 ± 0.046	0.362 ± 0.058
Micro F1	**0.470** ± 0.059	0.444 ± 0.040	0.420 ± 0.054
Exact Match	0.202 ± 0.083	0.173 ± 0.057	**0.207** ± 0.048

Mean performances across six random seeds for all models on the Hummingbird (top) and GoEmotions (bottom) datasets. Standard deviations are indicated with '±' and the highest performances per metric are boldfaced.

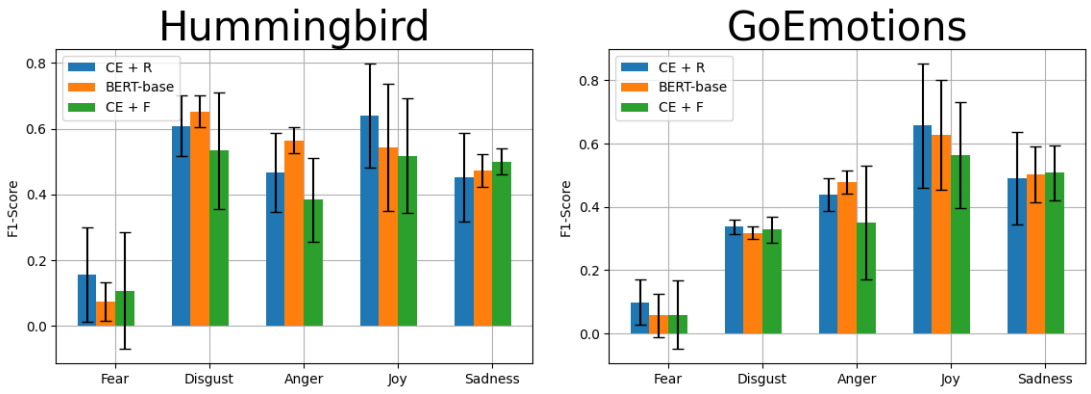


Figure 8: Class-wise performances for all models for emotion detection

Figure 8 provides a more nuanced insight into the ID and OOD performances of the models per class. Notably, CE + R and BERT-base perform similarly on each class individually on both datasets. However, CE + R on average achieves slightly higher performance than BERT base for three out of five classes on GoEmotions, namely 'Fear', 'Disgust' and 'Joy'. By performing the best on the majority of the classes, CE + R is able to

achieve the highest macro and micro F1 scores.

The performances of the three models on the GoEmotions dataset suggest that the components of attention regularisation contribute similarly to differences in OOD performance between CE + R and BERT-base. Specifically, the combination of L_{att} and the rationales $\{R_1, R_2, \ldots, R_N\}$ both contribute to a higher OOD performance than could be achieved by simply using L_{labels} or using L_{att} without the specific human identified features encoded in $\{R_1, R_2, \ldots, R_N\}$. However, the observed differences in performance across classes suggest that attention regularisation may positively impact the model's performance on some classes while potentially hindering its performance on others.

4.2.2 Learning Curves

Although CE + R achieves the highest OOD performance (in terms of macro and micro F1) when trained on the full dataset of 300 examples, these effects are not seen when smaller training sets are used. The learning curves in Figure 9 indicate that BERT-base is the first model to see improvements, clearly outperforming the other two models when trained on 100 examples, while CE + R performs much worse than the other two models. As more training data is added, CE + R and CE + F also begin to improve. When trained on the full training set of 300 texts CE + R and BERT-base overtake CE + F and reach a similar macro F1. While unexpected, these results suggest that attention regularisation is less beneficial in few-shot scenarios for emotion detection compared to sentiment analysis, possibly requiring more training data for these benefits to appear.

4.2.3 Similarity between attention weights and rationales

Although the differences in performance between CE + R and BERT-base are relatively small, Figure 10 indicates that attention regularisation increases the similarity between attention scores and rationales. CE + R achieves the highest similarity across all three metrics and for every emotional category. Furthermore, cosine similarity displays less divergence from the other two metrics compared to the sentiment analysis experiment. The absence of the 'Only Rationales' baseline could be a potential reason, as it recorded the highest cosine similarity across all sentiment analysis datasets. While more test samples are required to concretise these findings, this analysis indicates that attention regularisation is a promising technique to encourage BERT to focus on similar cues that humans do when determining the emotion that is expressed through a text. However, it is important to note that this analysis only concerns performance on the Hummingbird test set, as the GoEmotions dataset does not contain rationales.

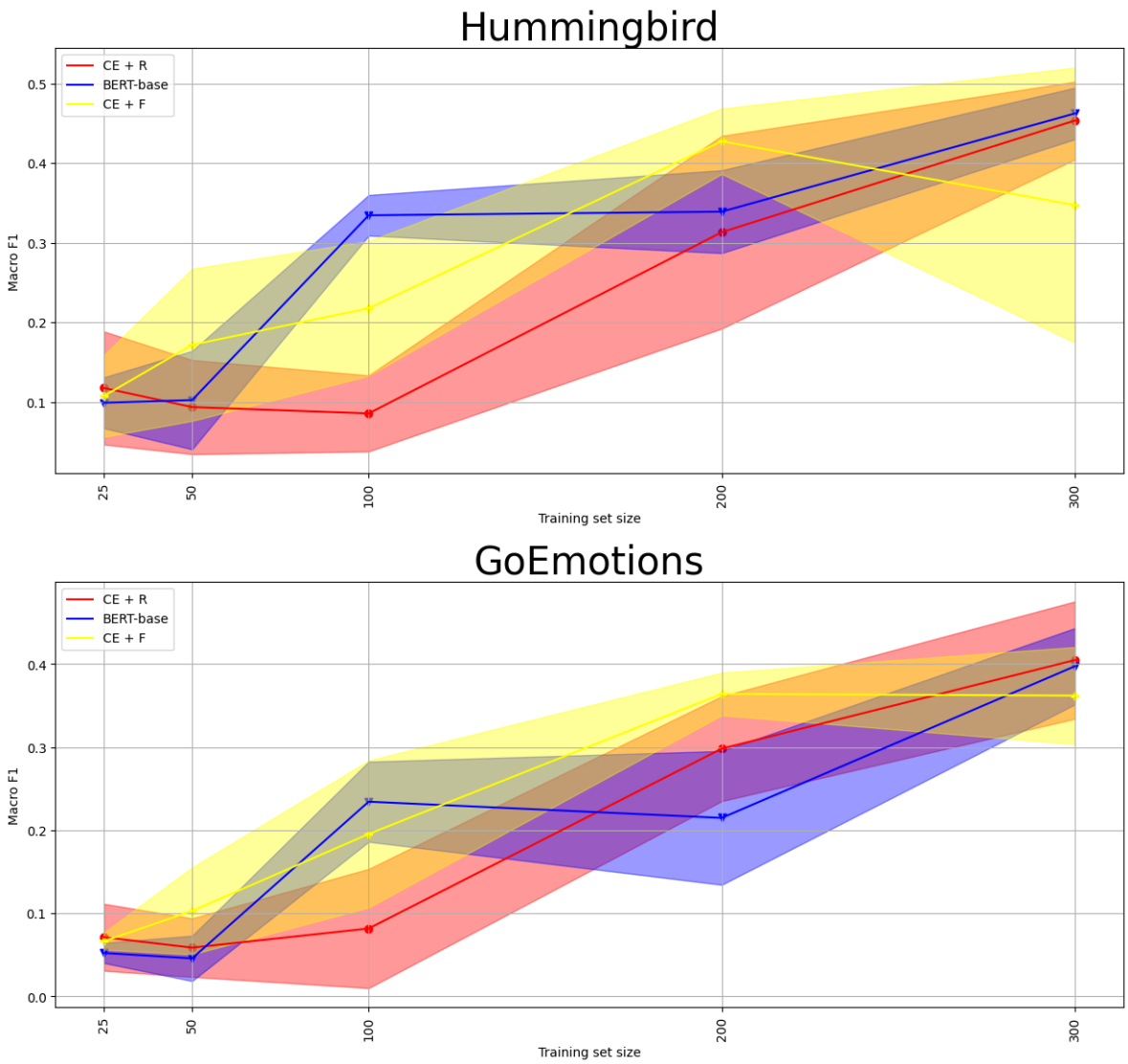


Figure 9: Learning curves for all models evaluated on the Hummingbird (top) and GoEmotions (bottom) datasets. Mean and standard deviations are plotted against the size of the training set used to train each model

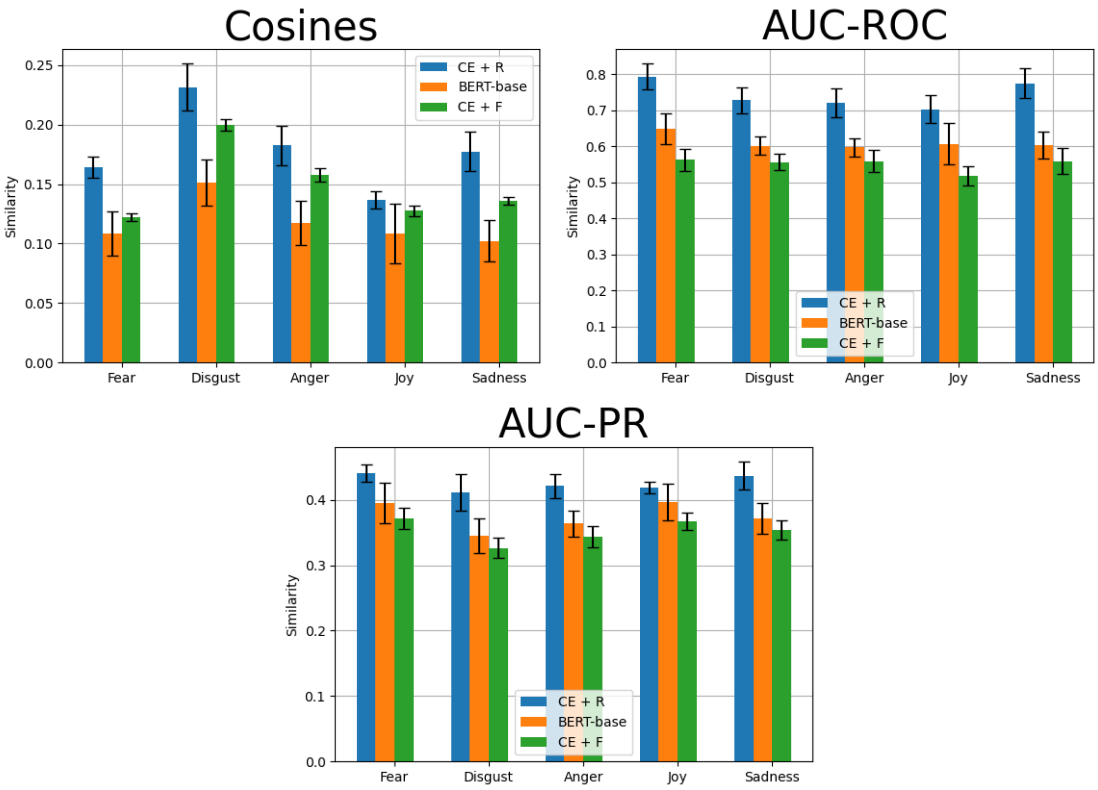


Figure 10: Cosine similarity (left), AUC-ROC (middle) and AUC-PR (right) of all three models per class

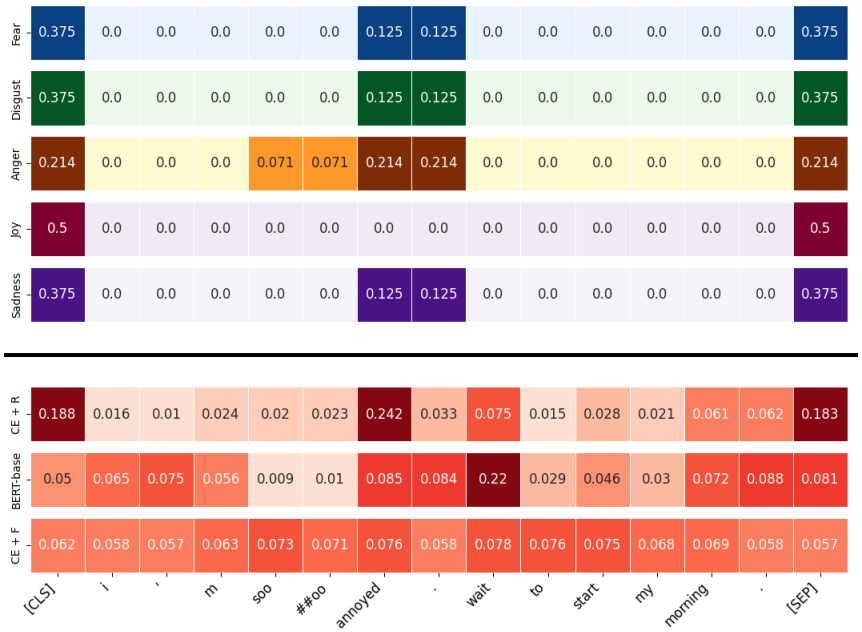
4.2.4 Qualitative analyses

The ability for attention scores and rationales to offer plausible explanations for a model's prediction, as observed in sentiment analysis, can differ across texts. Ideally, both texts examined in this subsection would originate from OOD datasets to better understand the relationship between similarity and performance in OOD settings. Yet, the GoEmotions dataset does not include rationales, making it impossible to determine whether attention regularisation assists BERT in focusing more on human-identified salient features. Therefore, an example from the Hummingbird dataset is explored first to enhance our understanding of this relationship.

Consider the following text from the Hummingbird test set:

i'm soooo annoyed. wait to start my morning.

Figure 11 displays the attention scores and rationale vectors for a single instance in the Hummingbird test set, providing insights into the explanation capabilities of CE + R and BERT-base models. The text is labeled as 'Anger', and CE + R accurately predicts this emotion by assigning the highest attention score to the token "annoyed," which aligns with the main salient feature identified by the rationale vector. However, both BERT-base and CE + F models fail to make any positive predictions, indicating their inability to classify the text into any emotional category. Examining the attention scores of BERT-base reveals a lack of clear focus on salient features, while CE + F distributes its attention evenly across all tokens. Consequently, no meaningful insights can be derived from the attention scores of these models in this particular case, suggesting their limited ability to leverage the information present in the text. Despite the apparent effectiveness of CE + R compared to the other models in this scenario, it is worth mentioning that despite the correct classification as 'Anger', CE + R also incorrectly assigns the text to the category of 'Disgust'. This example highlights the challenge of assessing the effectiveness of attention regularisation in improving model performance in multi-label classification settings, as models may still make incorrect predictions alongside correct ones.



	CE + R	BERT-base	CE + F
Cosine	**0.550**	0.296	0.488
AUC-ROC	**0.909**	0.682	0.727
AUC-PR	**0.417**	0.196	0.238

Figure 11: An example from the Hummingbird dataset were CE + R predicts the correct label and BERT-base does not. Tokens are highlighted according to the attention scores from the [CLS] token in the final layer. Rationale vectors (top) are separated from the attention scores of the models (bottom) by a straight line. Brighter colors indicate higher attention scores. Both attention scores and rationales are normalised to sum to one. The corresponding similarity scores for this example achieved by each model with the rationales for the correct class, namely 'Anger', are reported in the table underneath.

However, the interpretation of attention scores does not always provide an intuitive explanation for a model's decision. To illustrate this point, consider the following text:

Oops! The bottle was so oily it slipped outta my hand and into the trash!

Although this text is labeled as 'Sadness', it does not contain explicit cues that directly indicate sadness. The emotion of sadness is implied, as it can potentially be inferred from the writer's description of the bottle slipping out of their hand and ending up in the trash. The only emotion assigned to this review is the label 'Sadness'. However, the CE + R model, BERT-base, and CE + F incorrectly predict the emotions of 'Disgust' and 'Anger', while BERT-base and CE + F also correctly predict the label 'Sadness'. The fact that CE + F, which also employs attention regularisation, is able to make the correct classification suggests that the rationales used to regularise CE + R may be a contributing factor to its mistake. However, discerning the specific cause of this mistake from the attention scores depicted in Figure 12 is challenging. Similar to sentiment analysis, this example demonstrates that attention regularisation can sometimes work against the model's performance in certain scenarios.

The qualitative analyses conducted in this section present a nuanced perspective on the effectiveness of attention regularisation for emotion detection. On one hand, employing this approach can enhance the model's ability to make accurate predictions by identifying crucial patterns in the data that align with rationales provided by human annotators. However, due to the multi-label nature of emotion detection, correct classifications can coexist with incorrect ones. In cases where attention regularisation leads to a correct prediction but also introduces incorrect predictions, evaluating the efficacy of this technique becomes more intricate compared to binary classification settings. In such cases, it becomes essential to weigh the benefits of accurate predictions

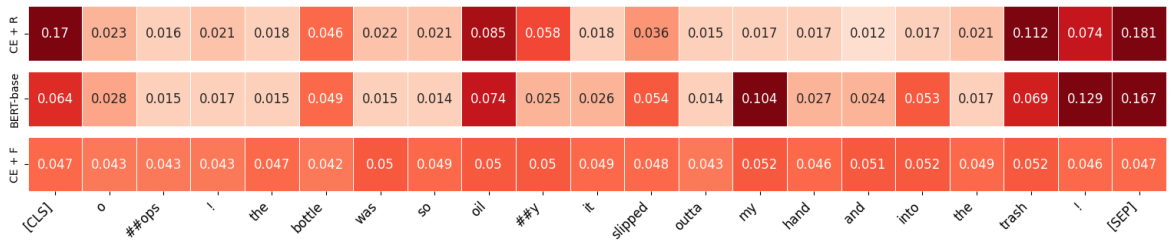


Figure 12: An example from the GoEmotions dataset were BERT-base makes the correct prediction and CE + R does not. The class label is 'Sadness'

against the drawbacks of inaccurate ones. Additionally, as exemplified in the second text, attention regularisation may even impede the model's capacity to correctly classify a text. Overall, while emotion detection encounters challenges akin to sentiment analysis, it introduces additional layers of complexity.

5 Discussion & Conclusion

5.1 Limitations

This study contains a number of limitations that ought to be considered by future work. These limitation can roughly grouped into five categories: (1) the models (2) the rationales, (3) the datasets for OOD evaluation, (4) the methods used to measure similarity and (5) the lack of significance testing. These five types of limitations are addressed in the following subsections.

5.1.1 Models

While BERT remains a popular model architecture for NLP experimenters, new and improved architectures are constantly being created that improve on the limitations of their predecessors. These improvements may nullify the benefits of attention regularisation by offering similar benefits. For example, DeBERTa uses a modified attention mechanism to capture both long-distance and local dependencies between words (He, Liu, Gao, & Chen, 2020). In a separate experiment, Stacey et al. (2022) observe that attention regularisation is less effective when combined with DeBERTa, yielding just 0.1% improvements in ID accuracy compared to 0.4% improvements in BERT[31]. Hence, modifications made to the architecture of pre-trained transformers, such as DeBERTa, may result in similar or the same improvements that attention regularisation can offer. However, more rigorous testing is required to make concrete conclusions on this matter.

The utility of the baselines constructed using the BERT architecture also requires some consideration. As outlined in Section 3.4.3, attention regularisation involves three key components: cross-entropy loss L_{labels} between the predicted and actual labels, L_{att}, and the rationale vectors $R_1, R_2, \ldots, R_N$ used to compute L_{att}. In this context, the CE + F baseline computes L_{att} using the 'flipped' versions of the rationales, namely by converting all positive entries into negative ones and vice versa. This baseline was designed to isolate the impact of L_{att} on OOD performance, regardless of the specific rationales used for its computation. However, CE + F uses rationale vectors with different proportions of positive values than CE + R to compute the term L_{att}, and thus may fail to fully isolate this term's effect as intended. This discrepancy is evidenced in Tables 11 and 12, which show the number of rationale and non-rationale words per text in the sentiment analysis and emotion detection experiments. This imbalance becomes especially noticeable for sentiment analysis, as each text in the IMDB dataset consists of just 10% rationales on average. Consequently, CE + F uses rationale vectors containing approximately nine times more positive values than CE + R to compute. Hence, it remains uncertain whether CE + R's superior OOD performance over CE + F is due to its exploitation of human insights encoded in the rationale vectors, or if CE + F's performance suffers from employing vectors to compute L_{att} whose positive values are distributed differently to those used by CE + R.

Although using state-of-the-art architectures and improved baselines may lead to different outcomes, these limitations do not undermine the findings of this study. Instead, they highlight the need for further experimentation to ascertain whether the observed patterns and conclusions hold across different models and architectures.

Statistic	IMDB	Yelp	BeerAdvocate
Mean Number of Rationales per Text	40.28	36.68	22.33
Mean Fraction of Rationales per Text	0.10	0.57	0.20
Ratio of Negative to Positive Rationales	1.62	0.95	0.91

Table 11: Rationale statistics for the IMDB, Yelp and BeerAdvocate datasets

5.1.2 Rationales

A limitation of using rationales from different datasets is that the ability of these rationales to effectively communicate salient features to the model may vary between these datasets. For instance, in the case of emotion detection, the rationales collected for the Hummingbird dataset by Hayati et al. (2021) contain aggregations of importance scores from multiple annotators, while the IMDB rationales collected by O. Zaidan et al. (2007)

[31] the OOD performance of DeBERTa is not reported

Statistic	Fear	Disgust	Anger	Joy	Sadness
Mean Number of Rationales per Text	4.41	6.89	6.79	4.38	5.544
Mean Fraction of Rationales per Text	0.09	0.17	0.16	0.96	0.13

Table 12: Rationale statistics for the Hummingbird dataset

were obtained from a single annotator. Additionally, identifying relevant words and phrases in longer texts like IMDB reviews is inherently more challenging, increasing the likelihood of overlooking important predictors of the class label. Consequently, the Hummingbird rationales may be more robust to the biases of a single annotator and potentially capture crucial portions of the text more effectively compared to the IMDB rationales. On the other hand, O. Zaidan et al. (2007) do perform a preliminary analysis that does indicate a degree of consistency in which words are highlighted by different annotators, which does verify the quality of their rationales to a degree. Consequently, the effectiveness of using rationales to guide the BERT's attention mechanism towards key features may be influenced the particular methods of rationale collection. Thus, the disparity in rationale quality across different datasets and the methods used for their collection can lead to inconsistent results between the sentiment analysis and emotion detection experiments, making it challenging to draw general conclusions about the efficacy of attention regularisation in enhancing OOD performance across different tasks.

Another important limitation of the rationales used in this study relates to their lack of heterogeneity. While aggregating over multiple annotators in the case of Hummingbird does create a degree of variation in the rationale vectors, these vectors still remain sparse. This sparsity can result in some words being overlooked that may still contribute to the overall meaning of the text despite not being explicitly highlighted. For example, in the sentiment analysis experiment, CE + F outperforms BERT-base and reaches comparable performance to CE + R by using all words that were not highlighted by annotators, suggesting that some words that are not highlighted can still contribute to successful predictions. Therefore, alternative approaches that use more granular methods to extract human attention may prove more effective. For example, Barrett et al. (2018) incorporate eye-tracking data into the fine-tuning process of an LSTM, which consistently outperforms baseline LSTMs on a number of NLP tasks, such as sentiment analysis, grammatical error detection and hate speech detection. In this way, alternative approaches that offer more granular methods for capturing human attention should be explored.

Overall, there is no clear consensus on how to curate rationales that offer the most benefits. Hence, more experimentation is encouraged to establish this consensus to further improve the predictive power of rationale-augmented models.

5.1.3 Datasets

The quality of rationales also depends on the quality of the texts they correspond to. While BeerAdvocate was included alongside Yelp as a more challenging OOD dataset, this dataset may also be sub-optimal for OOD evaluation. The BeerAdvocate dataset differs from the IMDB dataset not only in terms of input features but also in terms of class labels, as each label represents a specific aspect of beer rather than overall sentiment. In contrast, the Yelp and IMDB datasets share the same objective of providing a label that represents the overall meaning of the text, meaning Yelp and BeerAdvocate represent systematically different OOD scenarios, making direct comparisons between the two difficult when understand a model's OOD performance.

Furthermore, the emotion detection datasets present unique challenges in terms of label quality. Unlike sentiment analysis, where texts often come with explicit ratings from the original writers, emotion detection relies on annotators who are not the original writers, making it more difficult to provide accurate labels. This challenge is further amplified by the fact that datasets like Hummingbird and GoEmotions primarily consist of texts from social media platforms such as Twitter and Reddit, where the meaning of a text may only be understood within the context of a larger conversation. Given that annotators in both Hayati et al. (2021) and Demszky et al. (2020) are not provided with additional information about the texts they annotate, the quality of labels may suffer due to the lack of context. Therefore, there may be some scenarios in which attention regularisation successfully enables BERT to generalise from Hummingbird to GoEmotions, but an incorrect prediction is made due to the text being incorrectly labeled.

Aside from label quality, some of the sentiment analysis and emotion detection datasets may not be fully representative of a larger population of naturally occurring data. Specifically Yelp, reviews were removed for which either the majority label provided by the three annotators contradicts the gold label or no majority label exists. While doing so avoids confusion as to which rationales ought to be included for those reviews, it also results in potential edge cases being eliminated. Removing these edge cases makes the Yelp dataset 'easier' and thus results in a potential loss of diversity and real-world complexity, limiting the Yelp dataset's representativeness and applicability to more challenging scenarios.

Finally, it is important to consider their ecological validity, namely the degree to which the conditions set in the experimental setup truly resemble real-world scenarios, of the results of this experiment. To strengthen this validity, natural distribution shifts between the ID and OOD datasets were induced. However, the datasets used in this study raise concerns about their representativeness of real-world, unfiltered data due to specific criteria used for selection. For example, sentiment analysis datasets such as IMDB, Yelp, and BeerAdvocate only included reviews with clear positive or negative sentiments, excluding neutral reviews. This oversimplification of sentiment analysis may lead to confusion in real-world settings when classifying neutral reviews. Similarly BeerAdvocate, Hummingbird, and GoEmotions employed statistical measures or pilot classifiers to select texts where the desired class or emotion is most prominent, which may introduce sample selection bias by excluding texts that do not meet those criteria. Therefore, while it is sometimes unavoidable to have sample selection bias when curating datasets, it is crucial to acknowledge that the conclusions drawn from this thesis about the effects of attention regularisation may not necessarily apply to datasets that are more representative of a larger population of texts.

5.1.4 Methods used to measure similarity

It is important to consider the extent to which the similarity metrics used truly capture the overlap between rationales and attention scores. As discussed in Section 5.3.2, cosine similarity between attention scores and rationales does not directly translate to the attention scores predominantly being higher for words highlighted by humans. However, AUC-ROC and AUC-PR also have a number of limitations. Firstly, when using rationale vectors with continuous values, such as aggregations over multiple annotators or eye-tracking data, these metrics are less effective because these rationales must be binarised prior before a similarity score can be calculated, leading to a loss of information. Moreover, as discussed in Section 3.6, these metrics are ineffective if a text is assigned a rationale vector consisting of uniformly zeros (e.g. [0,0,0,0]), as there will be no true positives or false negatives. As a result, recall cannot be computed, leading to undefined values for AUC-ROC and AUC-PR. Similarly, if the rationale vector uniformly consists of ones (e.g. [1,1,1,1]), the false positive rate cannot be computed, resulting in an undefined AUC-ROC. In such cases, precision undeservedly becomes one for all threshold values due to the lack of false positives. Therefore, AUC-PR will not accurately reflect the tradeoff between precision and recall[32]. These issues were particularly relevant for the emotion detection experiment, where a substantial portion of the Hummingbird dataset had rationale vectors with uniform values per class. As these portions were omitted when calculating similarity scores, the findings of these analyses may not generalise to a larger population.

The justification for using similarity as a measure of alignment between BERT's attention mechanism and human-identified salient features deserves some scrutiny. In this scenario, attention scores form an explanation of the model's classification behavior and are used to estimate alignment with human behavior encapsulated in the rationales. However, the validity of these attention scores as suitable explanations can be contested. As DeYoung et al. (2020) points out, explanations derived from black-box models like BERT may be plausible but not necessarily faithful. In this study, the model explanations solely rely on attention scores from the [CLS] token from the final attention layer, excluding the majority of total computed attention scores. Hence, such explanations are unlikely to fully capture the overall behavior of the attention mechanism. A potential solution to this issue could involve adopting the approach of Jain and Wallace (2019), who suggest examining correlations between multiple explanation techniques— such as attention weights, gradient-based measures, and feature omission— to provide a more comprehensive picture.

[32]If precision is one for all threshold values, a straight line is created at the value one on the axis where precision is plotted. If precision is plotted on the y-axis, this results in a horizontal line at the top of the precision-recall graph, resulting in a rectangle with an area of one. Likewise, if precision is plotted on the x-axis, this results in a vertical line at the very right of the graph, meaning no area can be calculated.

5.1.5 Significance testing

Significance testing also plays a vital role in empirical research to ensure that the observed results are not due to chance. However, training models across multiple random seeds poses issues for significance testing that could not be addressed in this thesis. Comparing the performance of machine learning models is an ongoing area of research, and some studies (Dror, Baumer, Shlomov, & Reichart, 2018; Raschka, 2018; Dietterich, 1998) have suggested non-parametric tests such as McNemar's (McNemar, 1947) test for this purpose. However, when training models on multiple random seeds, conducting such tests becomes challenging due to the large number of comparisons involved. For instance, the sentiment analysis experiment would involve comparing four types of models with six random seeds each, resulting in 48 tests. While conducting these tests is straightforward, interpreting the outcomes is not. Alternatively, one can test whether differences between the models' mean performance across all seeds is statistically significant, but doing so must be done with care. For example, Stacey et al. (2022) used a two-tailed t-test to compare mean performances across 25 different seeds, they did not address whether the assumptions of this test were met. This thesis could have employed a non-parametric alternative, such as a Wilcoxon signed-rank test (Wilcoxon, 1947), for a more robust comparison, but using only six seeds means there is insufficient data to effectively conduct such a test. Overall, further exploring significance testing and adopting rigorous statistical methodologies are crucial for enhancing the reliability and meaningful interpretation of machine learning model comparisons.

5.2 Comparison to previous works

As stated in Section 1.4, this study strove to unify elements of previous works that tested the effectiveness of attention regularisation at improving the OOD performance of deep learning models on various tasks and domains. By comparing the results of this thesis with those of previous works, both for OOD (Section 5.2.1) and ID performance (5.2.2), this section provides an overview of how this thesis contributes to the larger body of research on rationale augmentation.

5.2.1 OOD performance

The results of the sentiment analysis experiment support those of previous works. Stacey et al. (2022), whose attention regularisation approach was adopted for this thesis, report improvements of ∼1% on an NLI task on two separate OOD datasets compared to a BERT model to which attention regularisation was not applied. These results are similar to those observed in the sentiment analysis experiment of this thesis, with improvements of ∼1% and ∼0.6% in accuracy being found on the Yelp and BeerAdvocate datasets respectively. These findings, alongside improvements in OOD performance on the emotion detection task, suggest that the attention regularisation approach used by Stacey et al. (2022) is robust across a number of different text classification tasks.

A number of important insights can further be gained by comparing this thesis with Bao et al. (2018), as one of their rationale-augmented models is similar to CE + R, but uses an LSTM architecture. This model only outperformed the same model that did not use rationales, suggesting that the OOD benefits of attention regularisation depend on the dataset used for evaluation. In my own results, CE + R outperforms BERT-base for both OOD datasets. However, many factors could be responsible for this difference between the two studies, such as model architecture, datasets and experimental setup. Regardless, the learning curves plotted by Bao et al. (2018) show similar trends to my own, with their rationale augmentation approach being highly effective when less training data is available. Overall, the benefits of attention regularisation do generalise to a degree across different model architectures and datasets, with the benefits being most evident in few-shot scenarios.

5.2.2 ID performance

While the main focus of this thesis is OOD performance, a short discussion of ID performance is warranted. In this study, CE + R and BERT-base models achieve a mean accuracy of ∼84.6% and ∼82.8%, respectively. These results are similar to those reported by Pruthi et al. (2022), who observed accuracy scores of ∼84.0% and ∼81.1% for their corresponding models. The similar ID performances in both experiments serves as an important benchmark, instilling confidence in the reliability and robustness of the results. However, it is worth noting that the current study averaged the models' performances over six runs with different random seeds. In contrast, Pruthi et al. (2022) did not explicitly mention this aspect. As explained in Section 3.4.5, this averaging strategy

provides a more robust estimation of the models' performance and helps account for the potential influence of random initialisation[33].

Despite the similarities between this thesis and Pruthi et al. (2022), it is worth noting that some studies that have also trained rationale-augmented models with this dataset have reported up to 90% test accuracy on the IMDB dataset, even when trained on less data than the 1200 texts employed in this research(Melamud et al., 2019; Wang et al., 2022). The source of this discrepancy in performance, whether it is tied to specific rationale augmentation procedures or other factors such as the use of average pooling over the head-first truncation strategy of applied in this study, remains ambiguous. Regardless, these works do not undermine this study's significance, which primarily aimed to investigate the effects of attention regularisation rather than attaining state-of-the-art performance.

5.2.3 Integrating previous insights on attention regularisation in OOD scenarios

In general, the findings of the sentiment analysis experiment align with the broader trends observed in previous work, particularly in relation to the improved OOD performance of rationale-augmented models in low-resource environments. The strength of this study lies in its synthesis of key elements from the work of Stacey et al. (2022), Bao et al. (2018), and Pruthi et al. (2022). This unifying approach reveals the broad-based effectiveness of attention regularisation, demonstrating its robustness across various architectures, tasks, and domains, thus positioning it beyond the realm of niche improvements.

5.3 Summary of results

This section provides summarises the results detailed in section 4 and the extent to which these results address the research questions posed in Section 1.5. The following subsections address each of these questions in turn.

5.3.1 SQ1: Does attention regularisation affect the OOD performance of pre-trained transformers similarly for sentiment analysis and emotion detection?

The motivation behind this question was to use sentiment analysis and emotion detection as a case study to provide initial insights into whether attention regularisation offers benefits in more challenging NLP tasks than binary classification. Sentiment analysis, an extensively studied task in previous literature, was strategically selected for this study as an initial point of reference for the application of attention regularisation. Following this experiment, the emotion detection experiment sought to extend these insights into the domain of multi-label classification, with the aim of discerning the potential benefits of attention regularisation in more intricate scenarios.

Overall, the effects of attention regularisation are consistent across both sentiment analysis and emotion detection tasks. Notably, CE + R achieves the highest OOD performance for all models on both tasks, reinforcing a common trend. However, as indicated in Sections 4.1.4 and 4.2.4, there exist scenarios where CE + R fails to to make correct classifications while BERT-base succeeds. These scenarios suggest that attention regularisation can benefit both tasks in certain areas, while being a hindrance in others. Despite these shared tendencies, one must exercise caution when making direct comparisons between sentiment analysis and emotion detection. For one, the model selection criteria varied across the two experiments, focusing on macro F1 scores for emotion detection to ensure a balanced performance across diverse classes, while sentiment analysis models were chosen based on accuracy. Moreover, as Figure 8 illustrates, attention regularisation's impact may be class-specific, which introduces additional complications that are inapplicable to binary classification settings.

In summary, while the limited sample size in the Hummingbird dataset constrains the preciseness of conclusions about attention regularisation in multi-label settings, the benefits of this approach are evident in both sentiment analysis and emotion detection tasks. Moreover, as acknowledged in Section 5.2, the sentiment analysis results mirror the trends observed in previous works, placing this study within an established body of research on attention regularisation. Consequently, validating the sentiment analysis results lends the emotion detection experiment a degree of legitimacy, given the similar experimental conditions used across both tasks.

[33]While Pruthi et al. (2022) computed L_{att} using KL-divergence between the rationales and attention scores, the current study followed the approach proposed by Stacey et al. (2022), computing L_{att} using mean-squared-error (MSE). This methodological difference could also contribute to the slight variation in performance.

This legitimacy has critical implications for expanding the scope of attention regularisation into more complex NLP tasks, notably multi-label classification. Hence, the consolidation of prior findings and novel experimentation with multi-label classification positions this study as a key contribution to the existing literature, charting new territories in the application of attention regularisation.

5.3.2 SQ2: To what extent does attention regularisation guide the attention mechanism of pre-trained transformers to align with salient features highlighted by human annotators?

The goal of measuring this similarity is twofold: (1) to understand whether attention regularisation gives rise to more explainable models and (2) whether the prioritisation of human-highlighted features is associated with higher OOD performance. Regarding the first goal, some examples discussed in Sections 4.1.4 and 4.2.4 where CE + R made the correct prediction and the other models do not, CE + R also seems to prioritise features deemed important by human annotators. Hence, the attention scores used in these analyses can be used to justify the model's prediction. However, other examples in which CE + R makes an incorrect prediction, there is no straightforward interpretation of why it makes that prediction. Hence, as explained in Section 5.1.4, using the attention scores of the [CLS] token in the final layer cannot always produce reliable or faithful explanations. In high-risk applications, where explainability is paramount, it is arguably even more important to understand why a model makes a mistake, as such mistakes may have devastating consequences. Therefore, while attention regularisation does provide some indication of producing a more explainable model, these findings must be taken with a large grain of salt.

Furthermore, the Mann-Whitney U-test illustrated in Figure 6 addresses the second goal by suggesting that higher similarity is associated with higher OOD performance, especially in the case of CE + R. Therefore, training models to maximise the similarity of their attention scores to rationales may yield more OOD-robust models. However, a significant difference between the similarity scores associated with positive predictions and those associated with negative predictions does not necessarily imply a causal relationship between similarity an performance. Simply maximising similarity alone may not guarantee optimal results, as other factors could be at play in determining the model's OOD performance. Additionally, this significance test only serves as a case study, as it was only performed for two models (CE + R and BERT-base) on a single seed (24) and for a single dataset (Yelp). To validate the findings of this analysis and establish a clearer link between OOD performance and similarity, more rigorous tests are required.

Conclusively, while this study presents a promising relationship between the use of attention regularisation, model explainability, and OOD performance. Attention regularisation appears to enhance the alignment of model behaviour with rationales in some instances, while the positive association observed between higher similarity and higher OOD performance further underscores the possibility of refining the attention regularisation approach to produce even more robust classifiers in future experiments. Nevertheless, it is essential to remember that the evidence supporting this relationship is derived from a limited context, and that more rigorous statistical testing is necessary to confirm these preliminary insights.

5.3.3 SQ3: Is attention regularisation an effective method for reducing the amount of training data required to achieve a desirable level of OOD performance in pre-trained transformers?

This question sought to understand whether attention regularisation serves as a useful tool in practical scenarios where labelled data is scarce and/or hard to come by. Based on the results, the usefulness of attention regularisation in these scenarios may vary between tasks. The sentiment analysis experiment showcased clear improvements in OOD performance when attention regularisation was used, especially when the models were trained 200 examples or less. However, these benefits were not observed for emotion detection, with CE + R unable to compete with the other two models until trained on upwards of 200 texts. Therefore attention regularisation is likely less beneficial in low-resource environments concerning emotion detection. However, given the small number of points plotted on learning curves in Figure 9, is unclear what trajectory the models' performances may take when trained on larger datasets. As discussed in Section 5.3.1, a higher quality and quantity of data is required to determine whether the results from the emotion detection experiment experiment are anomalous rather than representative of a larger performance trend. In sum, while practitioners using binary classifiers in low-resource domains likely benefit from attention regularisation, its practical utility for multi-label classifiers requires more rigorous testing.

5.4 Future work

The limitations highlighted in Section 5.1 present opportunities for future work to drive forward research into the practical application and improvement of rationale-augmented models. These opportunities can roughly be categorised in three groups: (1) improved methods for collecting rationales (2) improved techniques for measuring alignment with human-identified salient features and (3) rationale augmentation use cases beyond text classification. These three groups are discussed in the following subsections.

5.4.1 Improved methods for collecting rationales

As discussed in Section 5.1.2, there is no overarching framework for collecting high-quality rationales. Sen et al. (2020) made some initial strides in this direction by experimenting with different annotation strategies and providing insights into how the quality and content of rationales may differ from person to person. These insights are valuable for a study dedicated to developing rationales that optimise the OOD performance of ML models. One approach to develop such rationales could involve classifying the highlighted words according to their structural roles in a sentence (e.g., subject, predicate, object), or their lexical categories (e.g., nouns, verbs, adjectives). An in-depth investigation into these categories and how they correlate with OOD performance when attention regularisation is applied could provide valuable insights into how rationales can more effectively instill inductive bias in a model that allows for more effective generalisation across domains. In turn, the experimental setup used for rationale collection could be further refined by prioritising categories that are most associated with improved OOD performance.

Scenarios involving multi-label classification require even more consideration, as multiple rationale vectors for the same texts may point to similar features. Such rationales may cause a model to conflate different classes. In the case of Hummingbird, four of the five classes, namely fear, disgust and anger, represent negative emotions, while the fifth, namely joy, is a positive emotion. Accordingly, the qualitative analysis conducted in Section 4.2.4 (refer to Figure 11) showed that CE + R correctly predicted the text as 'anger', but also incorrectly assigns the label 'disgust', while the other models predict neither. Hence, CE + R may have experienced a phenomenon during training makes it challenging to separate these emotions. Figure 13, which shows the Spearman's correlation of the rationales for each emotion with one another, suggests that the rationales of the negative emotions are positively correlated with each other, but negatively correlated with 'joy'. As such, future studies addressing the impact of attention regularisation on OOD performance in multi-label scenarios may benefit from investigating whether 'de-correlated' rationale vectors allow a model to distinguish between classes more effectively when evaluated on OOD data.

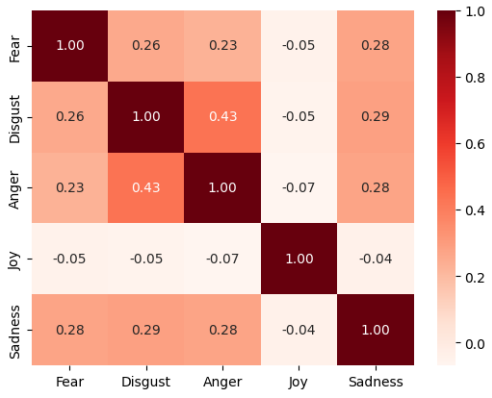


Figure 13: Spearman's correlation between the rationales for each class. All p-values were below 0.05.

Increasing the availability of rationales is equally important as, if not more important than increasing their quality. Currently, collecting rationales each time one aims to create a new rationale-augmented model for a specific task can be cumbersome and expensive. Exploring automated approaches to combat this issue may reduce the costs of collecting rationales from humans and allow for more efficient and systematic improvements to be made to models used in OOD scenarios. Works, such as Bao et al. (2018), Pruthi et al. (2022) and Hayati et al. (2023), have investigated approaches to create machine-generated rationales and their effectiveness in improving model

performance. For example, Pruthi et al. (2022) apply a 'teacher-student' approach whereby 'teacher models' were trained to perform sentiment analysis and question-answering tasks. Rationales were then extracted from the teacher model via different explanation techniques, such as attention-based, integrated gradients and layer conductance. New BERT and LSTM 'student models' were then augmented with these rationales and trained on a subset of the original dataset. These student models outperformed baseline models that were not augmented with these rationales. The authors did not directly compare these models to ones that were augmented with rationales taken from humans, nor did they investigate the benefits of the teacher-student approach in OOD scenarios. However, exploring these areas could shed light on whether machine-generated rationales offer similar insights to their human counterparts, or if additional experimentation is needed to do so.

5.4.2 Improved techniques for measuring alignment with human-identified salient features

Next to improving the quality and availability of rationales, it is also important to further refine the techniques used to align a model's behaviour with these rationales. As noted in Section 5.1.4, both the AUC-ROC and AUC-PR metrics are limited when the rationales contain continuous or uniform values. To address these issues, future experiments should incorporate a wider range of other similarity measures, such as Euclidean distance, Pearson's r, and Spearman's r. Furthermore, by employing multiple metrics, researchers can systematically investigate which metrics are most consistent with one another and whether this consistency actually reflects alignment with salient features. However these similarity metrics only measure the plausibility of a machine-generated explanations. Therefore, attention scores may not accurately reflect the inner workings of the model. Consequently, several studies have compared the faithfulness of various explanation techniques to address this concern (Nguyen, 2018; Atanasova, Simonsen, Lioma, & Augenstein, 2020; Jain & Wallace, 2019; Wiegreffe & Pinter, 2019; Jacovi & Goldberg, 2020; Serrano & Smith, 2019). By drawing from these studies, more authentic explanations can be crafted, thereby enabling similarity measures to produce outcomes that more reliably represent alignment with prominent features. By extension, one could explore using such explanations to modify attention regularisation and produce even more generalisable models.

5.4.3 Rationale augmentation use cases beyond text classification

Aside from refining rationale augmentation approaches, experimenting with these approaches outside of text classification settings may yield valuable insights and expand the applicability of rationale augmentation techniques. Given the rapid growth of NLP innovations in almost all sectors of society, the practical applications of rationale augmentation are endless. For example, rationales could be incorporated into existing approaches used to enable models to generate outputs that are more preferred by humans, such as the Reinforcement Learning through Human Feedback (RLHF) approach used to iterative fine-tune OpenAI's ChatGPT (Ouyang et al., 2022). Additionally, rationale augmentation may help improve fusion-based approaches that use attention mechanisms to map information from different media, such as text and images, to a common representation space (L. Dong et al., 2019; X. Liu, He, Chen, & Gao, 2019; Lu, Batra, Parikh, & Lee, 2019; Zhou et al., 2020). In this context, rationales could be included as part of the training signal to improve downstream tasks, such as image captioning and visual question answering. While these two examples represent just a fraction of the numerous potential applications for rationale augmentation approaches beyond text classification scenarios, they serve to illustrate these approaches' vast potential for expansion.

5.5 Conclusion

The goal of this thesis was to address the main research question stated in Section 1.5, namely

RQ: How does attention regularisation affect the OOD performance of pre-trained transformers?

This question was addressed from three distinct angles. Firstly, the influence of attention regularisation on OOD performance was evaluated in the context of both sentiment analysis and emotion detection. Despite their inherent differences, attention regularisation demonstrated potential for both tasks. In doing so, these experiments establish a baseline for future studies on rationale augmentation to go beyond experimenting with binary classification to more complex NLP tasks. The second approach provided a deeper understanding of how attention regularisation impacts the model's behaviour, particularly its tendency to highlight salient features identified by humans. The fact that CE + R demonstrated the highest overall similarity and OOD performance in both experiments provides some indication that attention regularisation is working as intended, namely to provide inductive bias that can be used to capture fundamental patterns in the training data rather than spurious

ones. These results provide initial steps to establishing a causal link between increased focus on salient features and increased OOD performance, which can be used to further refine rationale augmentation approaches and potentially create more explainable models. Lastly, attention regularisation's practical value was explored by assessing its OOD performance in low-resource environments. Here rationales were found to be notably beneficial for sentiment analysis. Although the same could not be said for emotion detection, this experiment is the first of its kind. As such, future developments and refinements may unlock benefits even in multi-label scenarios.

Studying the effects of attention regularisation on OOD performance from these three perspectives expands upon and unifies findings from previous works, such as Stacey et al. (2022), Pruthi et al. (2022) and Bao et al. (2018). Collectively, these perspectives contribute to a more comprehensive understanding of attention regularisation's effects on the OOD performance of pre-trained transformers, which can be used to create models that have a greater utility in practical scenarios. Furthermore, this thesis opens up a whole range of new possibilities, both regarding the refinement of rationale augmentation approaches as well as new real-world applications that may benefit from token-level supervision. As the field of machine learning continues to advance, the integration of rationales in various domains, such as generative and fusion-based tasks, has the potential to unlock new possibilities for performance improvement, explainability, and cross-modal advancements. With ongoing research and exploration, rationale augmentation can play a pivotal role in shaping the future of diverse NLP applications.

References

Abdou, M., Kulmizev, A., Hill, F., Low, D. M., & Søgaard, A. (2019, November). Higher-order comparisons of sentence encoder representations. 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)* (pp. 5838–5845). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-1593 doi: 10.18653/v1/D19-1593

Adhikari, A., Ram, A., Tang, R., & Lin, J. (2019). Docbert: Bert for document classification. *arXiv preprint arXiv:1904.08398*.

Amin, S., Neumann, G., Dunfield, K., Vechkaeva, A., Chapman, K. A., & Wixted, M. K. (2019). Mlt-dfki at clef ehealth 2019: Multi-label classification of icd-10 codes with bert. In *Clef (working notes)* (pp. 1–15).

Arjovsky, M. (2020). *Out of distribution generalization in machine learning* (Unpublished doctoral dissertation). New York University.

Atanasova, P., Simonsen, J. G., Lioma, C., & Augenstein, I. (2020, November). A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 3256–3274). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.emnlp-main.263 doi: 10.18653/v1/2020.emnlp-main.263

Bao, Y., Chang, S., Yu, M., & Barzilay, R. (2018, October-November). Deriving machine attention from human rationales. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 1903–1913). Brussels, Belgium: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D18-1216 doi: 10.18653/v1/D18-1216

Baroni, M. (2022). On the proper role of linguistically oriented deep net analysis in linguistic theorising. In *Algebraic structures in natural language* (pp. 1–16). CRC Press.

Barrett, M., Bingel, J., Hollenstein, N., Rei, M., & Søgaard, A. (2018). Sequence classification with human attention. In *Proceedings of the 22nd conference on computational natural language learning* (pp. 302–312).

Bau, A., Belinkov, Y., Sajjad, H., Durrani, N., Dalvi, F., & Glass, J. (2018). *Identifying and controlling important neurons in neural machine translation.*

Beltagy, I., Lo, K., & Cohan, A. (2019, November). SciBERT: A pretrained language model for scientific text. 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)* (pp. 3615–3620). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-1371 doi: 10.18653/v1/D19-1371

Betts, K. D., & Jaep, K. R. (2016). The dawn of fully automated contract drafting: Machine learning breathes new life into a decades-old promise. *Duke L. & Tech. Rev.*, *15*, 216.

Bhargava, P., Drozd, A., & Rogers, A. (2021, November). Generalization in NLI: Ways (not) to go beyond simple heuristics. In *Proceedings of the second workshop on insights from negative results in nlp* (pp. 125–135). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.insights-1.18 doi: 10.18653/v1/2021.insights-1.18

Bhat, M. M., Sordoni, A., & Mukherjee, S. (2021). Self-training with few-shot rationalization. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 10702–10712).

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, *29*.

Bordia, S., & Bowman, S. R. (2019, June). Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Student research workshop* (pp. 7–15). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N19-3002 doi: 10.18653/v1/N19-3002

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., ... others (2013). Api design for machine learning software: experiences from the scikit-learn project. *arXiv preprint arXiv:1309.0238*.

Canhoto, A. I., & Padmanabhan, Y. (2015). 'we (don't) know how you feel'–a comparative study of automated vs. manual analysis of social media conversations. *Journal of Marketing Management*, *31*(9-10), 1141–1157.

Carton, S., Kanoria, S., & Tan, C. (2022, May). What to learn, and how: Toward effective learning from rationales. In *Findings of the association for computational linguistics: Acl 2022* (pp. 1075–1088). Dublin, Ireland: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.findings-acl.86 doi: 10.18653/v1/2022.findings-acl.86

Chen, Q., Allot, A., Leaman, R., Islamaj, R., Du, J., Fang, L., ... others (2022). Multi-label classification for biomedical literature: an overview of the biocreative vii litcovid track for covid-19 literature topic annotations. *Database*, *2022*.

Clinchant, S., Jung, K. W., & Nikoulina, V. (2019, November). On the use of BERT for neural machine translation. In *Proceedings of the 3rd workshop on neural generation and translation* (pp. 108–117). Hong Kong: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-5611 doi: 10.18653/v1/D19-5611

Cui, R., Hershcovich, D., & Søgaard, A. (2022, July). Generalized quantifiers as a source of error in multilingual NLU benchmarks. In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 4875–4893). Seattle, United States: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.naacl-main.359 doi: 10.18653/v1/2022.naacl-main.359

Dale, R. (2019). Law and word order: Nlp in legal tech. *Natural Language Engineering*, *25*(1), 211–217.

Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., & Potts, C. (2013, 01). A computational approach to politeness with application to social factors. In (Vol. 1).

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the international aaai conference on web and social media* (Vol. 11, pp. 512–515).

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020, July). GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4040–4054). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.acl-main.372 doi: 10.18653/v1/2020.acl-main.372

Desai, S., & Durrett, G. (2020). Calibration of Pre-trained Transformers. In *Proceedings of the conference on empirical methods in natural language processing (emnlp).*

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N19-1423 doi: 10.18653/v1/N19-1423

DeYoung, J., Jain, S., Rajani, N. F., Lehman, E., Xiong, C., Socher, R., & Wallace, B. C. (2020, July). ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4443–4458). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.acl-main.408 doi: 10.18653/v1/2020

.acl-main.408

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, *10*(7), 1895–1923.

Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., ... Hon, H.-W. (2019). Unified language model pre-training for natural language understanding and generation. *Advances in neural information processing systems*, *32*.

Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science*, *14*, 241–258.

Dror, R., Baumer, G., Shlomov, S., & Reichart, R. (2018). The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1383–1392).

Eberle, O., Brandl, S., Pilot, J., & Søgaard, A. (2022). Do transformer models show similar attention patterns to task-specific human gaze? In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 4295–4309).

Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., & Hovy, E. (2021, August). A survey of data augmentation approaches for NLP. In *Findings of the association for computational linguistics: Acl-ijcnlp 2021* (pp. 968–988). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.findings-acl.84 doi: 10.18653/v1/2021.findings-acl.84

Fisch, A., Talmor, A., Jia, R., Seo, M., Choi, E., & Chen, D. (2019, November). MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd workshop on machine reading for question answering* (pp. 1–13). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-5801 doi: 10.18653/v1/D19-5801

Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. *Advances in neural information processing systems*, *31*.

Glockner, M., Shwartz, V., & Goldberg, Y. (2018, July). Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 650–655). Melbourne, Australia: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P18-2103 doi: 10.18653/v1/P18-2103

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S., & Smith, N. A. (2018, June). Annotation artifacts in natural language inference data. In *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 2 (short papers)* (pp. 107–112). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N18-2017 doi: 10.18653/v1/N18-2017

Guthier, B., Alharthi, R., Abaalkhail, R., & El Saddik, A. (2014). Detection and visualization of emotions in an affect-aware city. In *Proceedings of the 1st international workshop on emerging multimedia applications and services for smart cities* (pp. 23–28).

Hair Jr, J. F., & Sarstedt, M. (2021). Data, measurement, and causal inferences in machine learning: opportunities and challenges for marketing. *Journal of Marketing Theory and Practice*, *29*(1), 65–77.

Hartmann, M., & Sonntag, D. (2022, May). A survey on improving NLP models with human explanations. In *Proceedings of the first workshop on learning with natural language supervision* (pp. 40–47). Dublin, Ireland: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.lnls-1.5 doi: 10.18653/v1/2022.lnls-1.5

Hayati, S. A., Kang, D., & Ungar, L. (2021, November). Does BERT learn as humans perceive? understanding linguistic styles through lexica. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 6323–6331). Online and Punta Cana, Dominican Republic: Association for

Computational Linguistics. Retrieved from https://aclanthology.org/2021.emnlp-main.510 doi: 10.18653/v1/2021.emnlp-main.510

Hayati, S. A., Park, K., Rajagopal, D., Ungar, L., & Kang, D. (2023, May). StyLEx: Explaining style using human lexical annotations. In *Proceedings of the 17th conference of the european chapter of the association for computational linguistics* (pp. 2843–2856). Dubrovnik, Croatia: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.eacl-main.208

He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., ... others (2021). The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 8340–8349).

Hendrycks, D., Liu, X., Wallace, E., Dziedzic, A., Krishnan, R., & Song, D. (2020, July). Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 2744–2751). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.acl-main.244 doi: 10.18653/v1/2020.acl-main.244

Herrewijnen, E., Nguyen, D., Mense, J., Bex, F., et al. (2021). Machine-annotated rationales: Faithfully explaining text classification. In *35th aaai conference on artificial intelligence*.

Hertel, L., Collado, J., Sadowski, P., & Baldi, P. (2018). Sherpa: Hyperparameter optimization for machine learning models.

Hewitt, J., & Manning, C. D. (2019). A structural probe for finding syntax in word representations. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4129–4138).

Hollenstein, N., de la Torre, A., Langer, N., & Zhang, C. (2019, November). CogniVal: A framework for cognitive word embedding evaluation. In *Proceedings of the 23rd conference on computational natural language learning (conll)* (pp. 538–549). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/K19-1050 doi: 10.18653/v1/K19-1050

Hupkes, D., Dankers, V., Mul, M., & Bruni, E. (2020). Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, *67*, 757–795.

Hupkes, D., Giulianelli, M., Dankers, V., Artetxe, M., Elazar, Y., Pimentel, T., ... others (2022). State-of-the-art generalisation research in nlp: a taxonomy and review. *arXiv preprint arXiv:2210.03050*.

Jacovi, A., & Goldberg, Y. (2020, July). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4198–4205). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.acl-main.386 doi: 10.18653/v1/2020.acl-main.386

Jain, S., & Wallace, B. C. (2019, June). Attention is not Explanation. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 3543–3556). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N19-1357 doi: 10.18653/v1/N19-1357

Kanchinadam, T., Westpfahl, K., You, Q., & Fung, G. (2020). Rationale-based human-in-the-loop via supervised attention. In *Dash@ kdd*.

Karlgren, J., Sahlgren, M., Olsson, F., Espinoza, F., & Hamfors, O. (2012). Usefulness of sentiment analysis. In *Advances in information retrieval: 34th european conference on ir research, ecir 2012, barcelona, spain, april 1-5, 2012. proceedings 34* (pp. 426–435).

Kavumba, P., Takahashi, R., & Oda, Y. (2022). Are prompt-based models clueless? *Journal of Natural*

Language Processing, *29*(3), 991–996.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., ... others (2021). Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning* (pp. 5637–5664).

Lake, B., & Baroni, M. (2018). Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning* (pp. 2873–2882).

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). Albert: A lite bert for self-supervised learning of language representations. In *Iclr*. OpenReview.net. Retrieved from `http://dblp.uni-trier.de/db/conf/iclr/iclr2020.html#LanCGGSS20`

Lauriola, I., Lavelli, A., & Aiolli, F. (2022). An introduction to deep learning in natural language processing: Models, techniques, and tools. *Neurocomputing*, *470*, 443–456.

Lazaridou, A., Kuncoro, A., Gribovskaya, E., Agrawal, D., Liska, A., Terzi, T., ... others (2021). Mind the gap: Assessing temporal generalization in neural language models. *Advances in Neural Information Processing Systems*, *34*, 29348–29363.

Le Bras, R., Swayamdipta, S., Bhagavatula, C., Zellers, R., Peters, M., Sabharwal, A., & Choi, Y. (2020). Adversarial filters of dataset biases. In *International conference on machine learning* (pp. 1078–1088).

Lei, T., Barzilay, R., & Jaakkola, T. (2016, November). Rationalizing neural predictions. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 107–117). Austin, Texas: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/D16-1011` doi: 10.18653/v1/D16-1011

Liao, T., Taori, R., Raji, I. D., & Schmidt, L. (2021). Are we learning yet? a meta review of evaluation failures across machine learning. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (round 2)*.

Liu, X., Cheng, H., He, P., Chen, W., Wang, Y., Poon, H., & Gao, J. (2020). Adversarial training for large neural language models. *arXiv preprint arXiv:2004.08994*.

Liu, X., He, P., Chen, W., & Gao, J. (2019, July). Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 4487–4496). Florence, Italy: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/P19-1441` doi: 10.18653/v1/P19-1441

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019, Jul). Roberta: A robustly optimized bert pretraining approach. *Cornell University - arXiv*. doi: 10.48550/arxiv.1907.11692

Liusie, A., Raina, V., Raina, V., & Gales, M. (2022). Analyzing biases to spurious correlations in text classification tasks. In *Proceedings of the 2nd conference of the asia-pacific chapter of the association for computational linguistics and the 12th international joint conference on natural language processing* (pp. 78–84).

Lovering, C., Jha, R., Linzen, T., & Pavlick, E. (2021). Predicting inductive biases of pre-trained models. In *International conference on learning representations*.

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, *32*.

Lu, J., Yang, L., Namee, B., & Zhang, Y. (2022, May). A rationale-centric framework for human-in-the-loop machine learning. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 6986–6996). Dublin, Ireland: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2022.acl-long.481` doi: 10.18653/v1/2022

.acl-long.481

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 50–60.

Marcus, G. (2018). Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*.

Marcus, G. F. (1999). Connectionism: with or without rules?: Response to jl mcclelland and dc plaut (1999). *Trends in Cognitive Sciences*, *3*(5), 168–170.

Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., & Mukherjee, A. (2021). Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 35, pp. 14867–14875).

McAuley, J., Leskovec, J., & Jurafsky, D. (2012). Learning attitudes and attributes from multi-aspect reviews. In *2012 ieee 12th international conference on data mining* (pp. 1020–1025).

McCoy, T., Pavlick, E., & Linzen, T. (2019, July). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 3428–3448). Florence, Italy: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/P19-1334` doi: 10.18653/v1/P19-1334

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*(2), 153–157.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, *5*(4), 1093–1113.

Melamud, O., Bornea, M., & Barker, K. (2019). Combining unsupervised pre-training and annotator rationales to improve low-shot text classification. 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)* (pp. 3884–3893).

Miller, J., Krauth, K., Recht, B., & Schmidt, L. (2020). The effect of natural distribution shift on question answering models. In *International conference on machine learning* (pp. 6905–6916).

Min, B., Ross, H., Sulem, E., Veyseh, A., Nguyen, T., Sainz, O., ... Roth, D. (2021). *Recent advances in natural language processing via large pre-trained language models: A survey* (WorkingPaper). doi: https://doi.org/10.48550/arXiv.2111.01243

Mohammad, S., Bravo-Marquez, F., Salameh, M., & Kiritchenko, S. (2018). Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation* (pp. 1–17).

Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., & Herrera, F. (2012). A unifying view on dataset shift in classification. *Pattern recognition*, *45*(1), 521–530.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. The MIT Press.

Mutasodirin, M. A., & Prasojo, R. E. (2021). Investigating text shortening strategy in bert: Truncation vs summarization. In *2021 international conference on advanced computer science and information systems (icacsis)* (pp. 1–5).

Nam, J., Cha, H., Ahn, S., Lee, J., & Shin, J. (2020). Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, *33*, 20673–20684.

Nguyen, D. (2018). Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (pp. 1069–1078).

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., & Kiela, D. (2020, July). Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4885–4901). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.acl-main.441 doi: 10.18653/v1/2020.acl-main.441

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... others (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Pang, B., & Lee, L. (2005, June). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL'05)* (pp. 115–124). Ann Arbor, Michigan: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P05-1015 doi: 10.3115/1219840.1219855

Pang, B., Lee, L., & Vaithyanathan, S. (2002, July). Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 conference on empirical methods in natural language processing (EMNLP 2002)* (pp. 79–86). Association for Computational Linguistics. Retrieved from https://aclanthology.org/W02-1011 doi: 10.3115/1118693.1118704

Park, M., Cha, C., & Cha, M. (2012). Depressive moods of users portrayed in twitter. In *Proceedings of the 18th acm international conference on knowledge discovery and data mining, sigkdd 2012* (pp. 1–8).

Park, S. H., & Han, K. (2018). Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology*, *286*(3), 800–809.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... others (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, *32*.

Pham, H., Dai, Z., Ghiasi, G., Liu, H., Yu, A. W., Luong, M.-T., ... Le, Q. V. (2021). Combined scaling for zero-shot transfer learning. *arXiv preprint arXiv:2111.10050*.

Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R., & Van Durme, B. (2018, June). Hypothesis only baselines in natural language inference. In *Proceedings of the seventh joint conference on lexical and computational semantics* (pp. 180–191). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from https://aclanthology.org/S18-2023 doi: 10.18653/v1/S18-2023

Pruthi, D., Bansal, R., Dhingra, B., Soares, L. B., Collins, M., Lipton, Z. C., ... Cohen, W. W. (2022). Evaluating explanations: How much do explanations from the teacher aid students? *Transactions of the Association for Computational Linguistics*, *10*, 359–375.

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, *63*(10), 1872–1897.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... others (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training.

Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). Ai in health and medicine. *Nature medicine*, *28*(1), 31–38.

Rambocas, M., & Pacheco, B. G. (2018). Online sentiment analysis in marketing research: a review. *Journal of Research in Interactive Marketing*.

Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*.

Raunak, V., Kumar, V., & Metze, F. (2019). On compositionality in neural machine translation. *arXiv preprint arXiv:1911.01497*.

Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., & Goldberg, Y. (2020, July). Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7237–7256). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.acl-main.647 doi: 10.18653/v1/2020.acl-main.647

Ravi, S., & Larochelle, H. (2016). Optimization as a model for few-shot learning. In *International conference on learning representations.*

Ross, A., Peters, M. E., & Marasović, A. (2022). Does self-rationalization improve robustness to spurious correlations? *arXiv preprint arXiv:2210.13575*.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Schick, T., & Schütze, H. (2021, November). Generating datasets with pretrained language models. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 6943–6951). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.emnlp-main.555 doi: 10.18653/v1/2021.emnlp-main.555

Sechidis, K., Tsoumakas, G., & Vlahavas, I. (2011). On the stratification of multi-label data. *Machine Learning and Knowledge Discovery in Databases*, 145–158.

Sen, C., Hartvigsen, T., Yin, B., Kong, X., & Rundensteiner, E. (2020). Human attention maps for text classification: Do humans and neural networks focus on the same words? In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4596–4608).

Serrano, S., & Smith, N. A. (2019, July). Is attention interpretable? In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 2931–2951). Florence, Italy: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P19-1282 doi: 10.18653/v1/P19-1282

Sharma, A., Miner, A. S., Atkins, D. C., & Althoff, T. (2020). A computational approach to understanding empathy expressed in text-based mental health support. In *Emnlp.*

Sharma, M., Zhuang, D., & Bilgic, M. (2015). Active learning with rationales for text classification. In *Proceedings of the 2015 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 441–451).

Shen, Y., Heacock, L., Elias, J., Hentel, K. D., Reig, B., Shih, G., & Moy, L. (2023). *Chatgpt and other large language models are double-edged swords.* Radiological Society of North America.

Shen, Z., Liu, J., He, Y., Zhang, X., Xu, R., Yu, H., & Cui, P. (2021). Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*.

Sheng, D., & Yuan, J. (2021). An efficient long chinese text sentiment analysis method using bert-based models with bigru. In *2021 ieee 24th international conference on computer supported cooperative work in design (cscwd)* (pp. 192–197).

Shum, H.-Y., He, X.-d., & Li, D. (2018). From eliza to xiaoice: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, *19*, 10–26.

Simon, H. A. (1990). Invariants of human behavior. *Annual review of psychology*, *41*(1), 1–20.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1631–1642).

Søgaard, A. (2016). Evaluating word embeddings with fmri and eye-tracking. In *Proceedings of the 1st workshop on evaluating vector-space representations for nlp* (pp. 116–121).

Sood, E., Tannert, S., Frassinelli, D., Bulling, A., & Vu, N. T. (2020, November). Interpreting attention models with human visual attention in machine reading comprehension. In *Proceedings of the 24th conference on computational natural language learning* (pp. 12–25). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.conll-1.2 doi: 10.18653/v1/2020.conll-1.2

Sovrano, F., Palmirani, M., & Vitali, F. (2022). Combining shallow and deep learning approaches against data scarcity in legal domains. *Government Information Quarterly*, *39*(3), 101715.

Stacey, J., Belinkov, Y., & Rei, M. (2022). Natural language inference with a human touch: Using human explanations to guide model attention. In *Proceedings of the thirty-sixth aaai conference on artificial intelligence (aaai 2022).*

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune bert for text classification? In *Chinese computational linguistics: 18th china national conference, ccl 2019, kunming, china, october 18–20, 2019, proceedings 18* (pp. 194–206).

Szymański, P., & Kajdanowicz, T. (2017, February). A scikit-based Python environment for performing multi-label classification. *ArXiv e-prints*.

Szymański, P., & Kajdanowicz, T. (2017). A network perspective on stratification of multi-label data. In L. Torgo, B. Krawczyk, P. Branco, & N. Moniz (Eds.), *Proceedings of the first international workshop on learning with imbalanced domains: Theory and applications* (Vol. 74, pp. 22–35). ECML-PKDD, Skopje, Macedonia: PMLR.

Talman, A., & Chatzikyriakidis, S. (2018). Testing the generalization power of neural network models across nli benchmarks. *arXiv preprint arXiv:1810.09774*.

Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., & Schmidt, L. (2020). Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, *33*, 18583–18599.

Tenney, I., Das, D., & Pavlick, E. (2019, July). BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 4593–4601). Florence, Italy: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P19-1452 doi: 10.18653/v1/P19-1452

Tu, L., Lalwani, G., Gella, S., & He, H. (2020). An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, *8*, 621–633.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wang, J., Sharma, M., & Bilgic, M. (2022). Ranking-constrained learning with rationales for text classification. In *Findings of the association for computational linguistics: Acl 2022* (pp. 2034–2046).

Warstadt, A., & Bowman, S. R. (2020). Can neural networks acquire a structural bias from raw linguistic data? *arXiv preprint arXiv:2007.06761*.

Wei, J., & Zou, K. (2019, November). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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)* (pp. 6382–6388). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-1670 doi: 10.18653/v1/D19-1670

Wiegreffe, S., & Pinter, Y. (2019, November). Attention is not not explanation. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint*

conference on natural language processing (emnlp-ijcnlp) (pp. 11–20). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-1002 doi: 10.18653/v1/D19-1002

Wilcoxon, F. (1947). Probability tables for individual comparisons by ranking methods. *Biometrics*, *3*(3), 119–122.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... others (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (pp. 38–45).

Wortsman, M., Ilharco, G., Kim, J. W., Li, M., Kornblith, S., Roelofs, R., ... others (2022). Robust fine-tuning of zero-shot models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 7959–7971).

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, *32*.

Yao, H., Chen, Y., Ye, Q., Jin, X., & Ren, X. (2021). Refining language models with compositional explanations. *Advances in Neural Information Processing Systems*, *34*, 8954–8967.

Yu, K.-H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature biomedical engineering*, *2*(10), 719–731.

Zahera, H. M., Elgendy, I. A., Jalota, R., Sherif, M. A., & Voorhees, E. (2019). Fine-tuned bert model for multi-label tweets classification. In *Trec* (pp. 1–7).

Zaidan, O., & Eisner, J. (2008). Modeling annotators: A generative approach to learning from annotator rationales. In *Proceedings of the 2008 conference on empirical methods in natural language processing* (pp. 31–40).

Zaidan, O., Eisner, J., & Piatko, C. (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* (pp. 260–267).

Zaidan, O. F., Eisner, J., & Piatko, C. (2008). Machine learning with annotator rationales to reduce annotation cost. In *Proceedings of the nips* 2008 workshop on cost sensitive learning* (pp. 260–267).

Zhang, D., Sen, C., Thadajarassiri, J., Hartvigsen, T., Kong, X., & Rundensteiner, E. (2021). Human-like explanation for text classification with limited attention supervision. In *2021 ieee international conference on big data (big data)* (pp. 957–967).

Zhang, X., Song, X., Feng, A., & Gao, Z. (2021). Multi-self-attention for aspect category detection and biomedical multilabel text classification with bert. *Mathematical Problems in Engineering*, *2021*, 1–6.

Zhang, Y., Baldridge, J., & He, L. (2019, June). PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 1298–1308). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N19-1131 doi: 10.18653/v1/N19-1131

Zhang, Y., Marshall, I., & Wallace, B. C. (2016). Rationale-augmented convolutional neural networks for text classification. In *Proceedings of the conference on empirical methods in natural language processing. conference on empirical methods in natural language processing* (Vol. 2016, p. 795).

Zhang, Y., & Yang, Q. (2018). An overview of multi-task learning. *National Science Review*, *5*(1), 30–43.

Zhang, Z., Liu, J., & Razavian, N. (2020, November). BERT-XML: Large scale automated ICD coding using

BERT pretraining. In *Proceedings of the 3rd clinical natural language processing workshop* (pp. 24–34). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.clinicalnlp-1.3 doi: 10.18653/v1/2020.clinicalnlp-1.3

Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., & Gao, J. (2020). Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 34, pp. 13041–13049).

Zhu, C., Cheng, Y., Gan, Z., Sun, S., Goldstein, T., & Liu, J. J. (2020, April). Freelb: Enhanced adversarial training for natural language understanding. In *Eighth international conference on learning representations (iclr)*. Retrieved from https://www.microsoft.com/en-us/research/publication/freelb-enhanced-adversarial-training-for-natural-language-understanding/

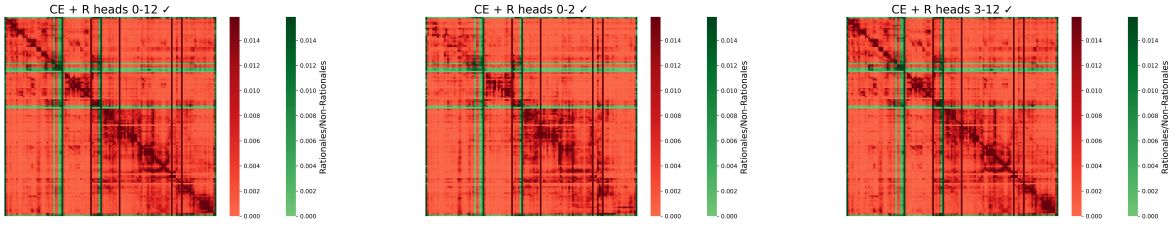
Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the ieee international conference on computer vision* (pp. 19–27).

Ziora, L. (2016). The sentiment analysis as a tool of business analytics in contemporary organizations. *Studia Ekonomiczne*, *281*, 234–241.

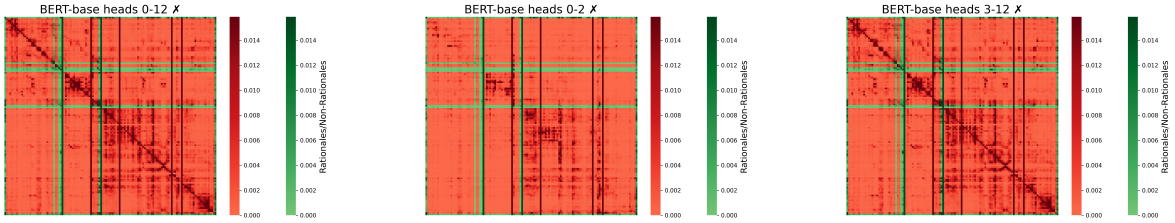
Zou, J., Zhang, Y., Jin, P., Luo, C., Pan, X., & Ding, N. (2021). Palrace: Reading comprehension dataset with human data and labeled rationales. *arXiv preprint arXiv:2106.12373*.

A Potential effects of attention regularisation on specific attention heads

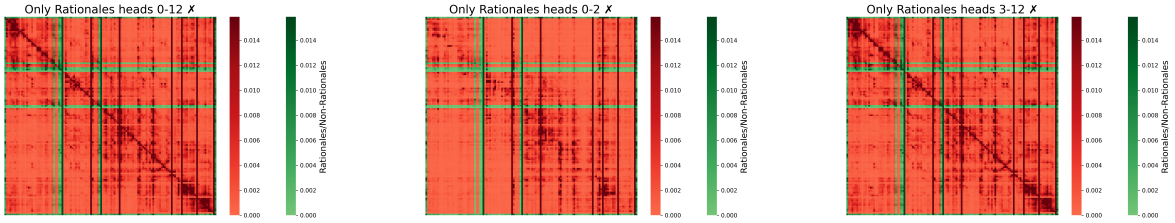
Attention regularisation is potentially an effective method for increasing activity in specific attention heads. Figure 14 shows heatmaps of the attention matrices of all four models for the sentiment analysis from the final attention layer. These heatmaps suggest that CE + R disperses its attention across more tokens compared to BERT-base and focuses more on the rationales, especially in the first two heads. In this way, attention regularisation may encourage specific attention heads, that would otherwise be less inclined, to focus on relevant features. Further experimentation with specific attention heads may yield interesting insights into how attention regularisation can be refined by directly eliciting specific behaviour from those heads.



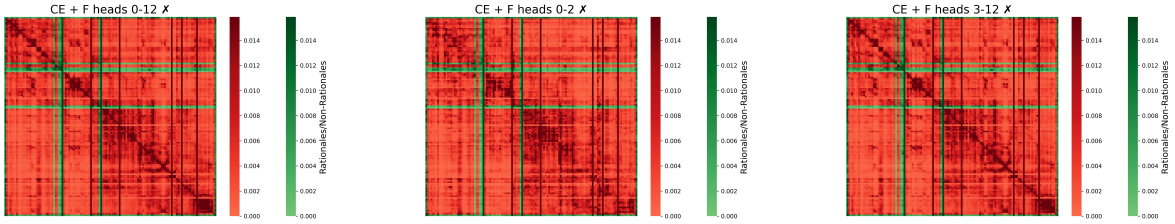
(a) CE + R



(b) BERT-base



(c) Only Rationales



(d) CE + F

Figure 14: A comparison of attention heatmaps of each model on one review in the Yelp dataset. Each model is represented by a heatmap of the attention matrix from the final layer averaged over all 12 attention heads (left), the first two attention heads (middle) and the last 9 attention heads (center). All attention scores to and from tokens highlighted by annotators are colored blue while all other attention scores are colored red. A checkmark (✓) indicates a correct prediction whle a cross (✗) indicates an incorrect prediction.

B Optimal hyperparameters found by each model

The optimal hyperparameter settings for the sentiment analysis and emotion detection experiment are displayed in Tables 13 and 14. Both experiments show opposite trends, with CE + R preferring a larger number of attention heads to compute L_{att} than CE + F, but a smaller value for λ. For emotion detection, the reverse is true.

Hyperparameter	CE + R	BERT-base	Only Rationales	CE + F
Attention Heads	10	NA	NA	2
Lambda	0.6	NA	NA	1.4
Learning Rate	1e-5	1e-5	5e-5	1e-5

Table 13: Optimal hyperparameter settings for all models for the sentiment analysis task

Hyperparameter	CE + R	BERT-base	CE + F
Attention Heads	2	NA	6
Lambda	1.4	NA	0.6
Learning Rate	7.5e-5	7.5e-5	5e-5

Table 14: Optimal hyperparameter settings for all models for the emotion detection task

C Correlation between similarity metrics

While AUC-ROC and AUC-PR are commonly favored metrics, the analysis presented in Table 15c reveals a strong correlation between AUC-PR and cosine similarity for CE + R, in contrast to the correlation between AUC-PR and AUC-ROC. This discrepancy between the different similarity measures raises doubts about the extent to which CE + R truly utilises rationales to make predictions in OOD settings. Consequently, further investigations and analyses are necessary to better understand the relationship between these similarity measures and their alignment with human-identified salient features.

Metric	Cosines	AUC-ROC	AUC-PR
Cosines	1.0 ± 0.0	0.1597 ± 0.039	0.6719 ± 0.0168
AUC-ROC	0.1597 ± 0.039	1.0 ± 0.0	0.1353 ± 0.0243
AUC-PR	0.6719 ± 0.0168	0.1353 ± 0.0243	1.0 ± 0.0

(a) Yelp

Metric	Cosines	AUC-ROC	AUC-PR
Cosines	1.0 ± 0.0	0.3093 ± 0.0365	0.6605 ± 0.0231
AUC-ROC	0.3093 ± 0.0365	1.0 ± 0.0	0.2942 ± 0.0321
AUC-PR	0.6605 ± 0.0231	0.2942 ± 0.0321	1.0 ± 0.0

(b) BeerAdvocate

(c) Spearman's correlation between all similarity metrics of CE + R on the Yelp (top) and BeerAdvocate (bottom) datasets. All coefficients are averaged across all six random seeds and standard deviations are indicated with '±'. All p-values are smaller than 0.05.