Modeling Complex Interactions in Long Documents for Aspect-Based Sentiment Analysis

Zehong Yan[1], Wynne Hsu[1], Mong Li Lee[1], David Roy Bartram-Shaw[2]

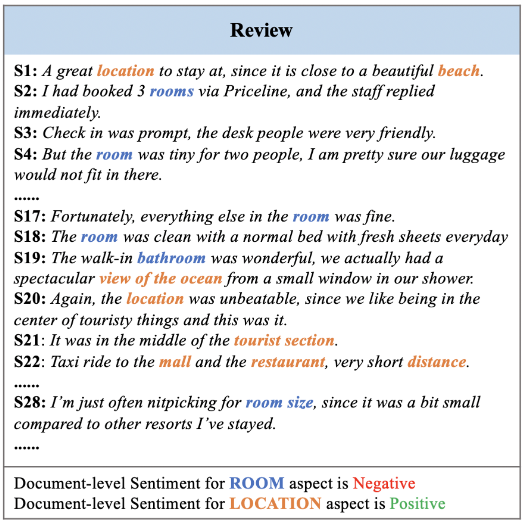
[1]NUS Centre for Trusted Internet & Community, National University of Singapore
[2]Edelman Data & Intelligence
{zehong, whsu, leeml}@comp.nus.edu.sg, david.bartram.shaw@gmail.com
https://yanzehong.github.io/dart/

Abstract

The growing number of online articles and reviews necessitates innovative techniques for document-level aspect-based sentiment analysis. Capturing the context in which an aspect is mentioned is crucial. Existing models have focused on relatively short reviews and may fail to consider distant contextual information. This is especially so in longer documents where an aspect may be referred to in multiple ways across dispersed sentences. This work introduces a hierarchical Transformer-based architecture that encodes information at different level of granularities with attention aggregation mechanisms to learn the local and global aspect-specific document representations. For empirical validation, we curate two datasets of long documents: one on social issues, and another covering various topics involving trust-related issues. Experimental results show that the proposed architecture outperforms state-of-the-art methods for document-level aspect-based sentiment classification. We also demonstrate the potential applicability of our approach for long document trust prediction.

1 Introduction

As user-generated content on the web continues to multiply at an exponential rate, the need for automated sentiment in these documents has grown markedly. The ability to discover fine-grained sentiments can provide valuable insights as to how, why, and where an entity is liked and trusted[1]. Early works have focused on classifying the overall sentiment of a document (Yang et al., 2016; Turney, 2002; Diao et al., 2023), while subsequent research performs aspect-based sentiment analysis to identify the fine-grained sentiments concerning the different aspects of some target entity (Severyn and Moschitti, 2015; Pontiki et al., 2016; Nazir et al., 2020; Brauwers and Frasincar, 2021).



Review

S1: *A great **location** to stay at, since it is close to a beautiful **beach**.*
S2: *I had booked 3 **rooms** via Priceline, and the staff replied immediately.*
S3: *Check in was prompt, the desk people were very friendly.*
S4: *But the **room** was tiny for two people, I am pretty sure our luggage would not fit in there.*
......
S17: *Fortunately, everything else in the **room** was fine.*
S18: *The **room** was clean with a normal bed with fresh sheets everyday*
S19: *The walk-in **bathroom** was wonderful, we actually had a spectacular **view of the ocean** from a small window in our shower.*
S20: *Again, the **location** was unbeatable, since we like being in the center of touristy things and this was it.*
S21: *It was in the middle of the **tourist section**.*
S22: *Taxi ride to the **mall** and the **restaurant**, very short **distance**.*
......
S28: *I'm just often nitpicking for **room size**, since it was a bit small compared to other resorts I've stayed.*
......

Document-level Sentiment for **ROOM** aspect is Negative
Document-level Sentiment for **LOCATION** aspect is Positive

Figure 1: Sample hotel review.

Aspect-based sentiment analysis can be performed at the sentence-level or document-level. Sentence-level aspect-based sentiment analysis focuses on independently classifying the sentiments associated with aspects in individual sentences (Peng et al., 2020; Yan et al., 2021). However, this approach fails to consider the context of the aspect, which can often be inferred from preceding or succeeding sentences or paragraphs. In Figure 1, the sentiment expressed toward the aspect "Location" is not clear just by looking at sentence **S21**. By examining the surrounding sentences **S1**, **S20** and **S22**, which are all positive, one could infer that the phrase "in the middle of the tourist section" has a positive sentiment, demonstrating the importance of *context* in aspect-based sentiment analysis at the document level. Further, sentences in the same document may express *conflicting sentiments* towards the same aspect. For example, sentence **S17**, **S18** and **S19** express a positive sentiment towards the aspect "Room", but **S4** and **S28** convey a negative sentiment. Simply classifying the overall sentiment based on a single sentence or taking the majority vote may led to incorrect conclusions.

[1] https://www.edelmandxi.com/trust-intelligence/measuring-trust-prerequisite-unlocking-growth

Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 23–34
August 15, 2024 ©2024 Association for Computational Linguistics

In this work, we design a hierarchical Transformer-based architecture called DART that leverages multiple layers of attention mechanisms. This allows us to capture the dependencies among sentences in long documents and learn aspect-specific document representations. DART performs attention aggregation on the learned representations to take into account both the local and global contexts. By employing learnable global aspect queries, our model aggregates sentiments that reflects the overall sentiment of the document, even in the presence of conflicting sentiments.

We curate two datasets, one focusing on social issues and another on trust-related issues. Initial experiments indicate that even GPT-4 has difficulty dealing with implicit aspects and often misinterprets sentiment due to insufficient aspect knowledge. Comprehensive experiments show that DART achieves state-of-the-art accuracy for document-level aspect-based sentiment classification, and is also effective in predicting trust and polarity in long complex documents.

2 Related Work

Research on aspect-based sentiment analysis can be broadly classified into sentence level and document level. Sentence-level aspect-based sentiment analysis includes using Long Short-Term Memory (LSTM) network to model aspects in sentences (Tang et al., 2016), attention-based LSTM to correlate aspects and sentiment polarities (Wang et al., 2016; Ma et al., 2017; Tay et al., 2018), deep memory networks to integrate aspect information (Tang et al., 2016; Chen et al., 2017), and gated networks to select aspect-specific sentiment in sentences (Zhang et al., 2016; Xue and Li, 2018). (Chen et al., 2020) introduce graph attention networks to improve sentence prediction by incorporating sentiment preference information from the document context. The work in (Yan et al., 2021) propose a unified framework for fine-grained sentiment analysis to identify the aspect and opinion terms as well as its sentiment polarity for each sentence.

Document-level aspect-based sentiment analysis predicts the sentiment polarity for each aspect mentioned in a document. Traditional approaches have largely relied on feature engineering. Latent rating regression (LRR) (Wang et al., 2010) is a probabilistic graphical model that generates document sentiment representation from a weighted sum of the latent aspect variables. (Lu et al., 2011) use support vector regression model based on handcrafted features to predict aspect ratings. To handle correlation between aspects, (McAuley et al., 2012) add a dependency term that explicitly encodes relationships between aspects. These methods have strict assumptions about words and sentences such as whether a word is an aspect or sentiment towards an aspect, and typically use bag-of-words representations which are insufficient to capture the order of words and complex semantics.

Neural network methods for document-level aspect sentiment analysis include N-DMSC (Yin et al., 2017), VWS-DMSC (Zeng et al., 2019) and D-MILN (Ji et al., 2020). N-DMSC employs hierarchical LSTM to create aspect-aware document representations using question-answer pairs constructed from aspect-related keywords and aspect ratings. VWS-DMSC uses a multi-task learning framework with rules to extract target-opinion word pairs to guide the sentiment prediction towards document aspects in a weakly supervised manner. D-MILN is a multiple instance learning network that models the relation between aspect-level and document-level sentiment with document-level supervision. (Fei et al., 2021) model the latent target-opinion distribution as prior information and employ a two-layer BiLSTM to obtain the overall document-level sentiment classification.

Transformer models have been utilized for aspect sentiment analysis (Fei et al., 2022; Islam and Bhattacharya, 2022). However, they are limited to processing sequences of up to 512 tokens. To overcome this limitation, models such as Longformer (Beltagy et al., 2020), Big Bird (Zaheer et al., 2020), Hi-Transformer (Wu et al., 2021) and LongT5 (Guo et al., 2022) have been introduced. However, these models have not yet been specifically utilized for aspect-based sentiment analysis.

3 Proposed Framework

The proposed DART framework takes as input a document d and an aspect a_j and outputs the predicted sentiment for a_j. Figure 2 shows the architecture of DART which consists of four key blocks:

Sentence Encoding Block. This block focuses on transforming the document into individual sentences and using a pretrained language model to generate representations for every sentence-aspect combination.

Global Context Interaction Block. This block employs dual transformer encoders to model interac-

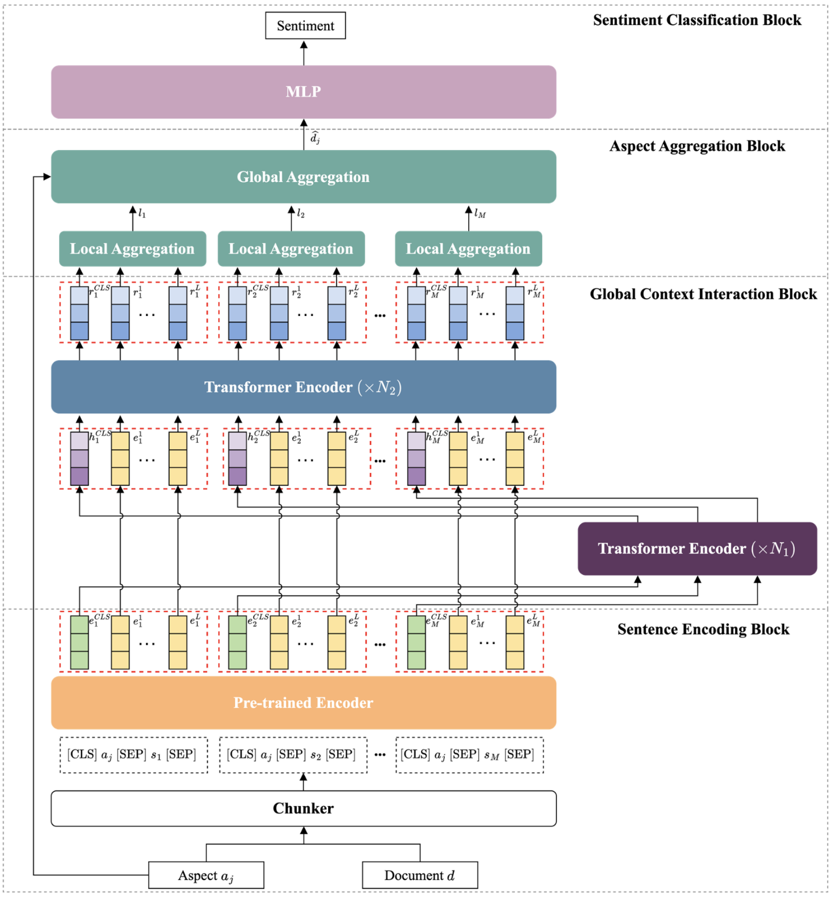


Figure 2: Overview of DART framework.

tions among sentences and generate context-aware sentence embeddings. This is a crucial component of DART as it captures essential aspect-specific information across long-range dependencies.

Aspect Aggregation Block. This block aggregates the contextually enriched sentence embedding to produce an aspect-specific representation of the entire document.

Sentiment Classification Block. With the document representation obtained, this block leverages a two-layered Multilayer Perceptron (MLP) to predict the sentiment for the aspect.

3.1 Sentence Encoding Block

Initially, the input document is divided into M sentences, denoted as $s_1, s_2, ..., s_M$. This is achieved using the sentence splitter from the Natural Language Toolkit[2]. Then we construct fixed length sequences $seq_1, seq_2, ...,$ seq_M, adding right paddings if needed. Each sequence seq_i is given by:

$$seq_i = \texttt{[CLS]}\ a_j\ \texttt{[SEP]}\ s_i\ \texttt{[SEP]}$$

where `[CLS]` and `[SEP]` are the special tokens to denote the sentence-level information and separator respectively. The sequence seq_i is fed into a BERT-based pretrained model to generate the embedding

$$[e_i^{CLS}, e_i^1, e_i^2, ..., e_i^L]$$

where e_i^k is the k^{th} token in seq_i and L is the fixed length of the sequence.

3.2 Global Context Interaction Block

This block captures dependencies among sentences so that a sentence can be understood in the broader context of the entire document, thus increasing the accuracy of sentiment prediction for a specific aspect. It incorporates two transformer encoders which serve different purposes.

The first transformer encoder focuses on the inter-sentence relationships. It uses the $[CLS]$ tokens which are condensed representations of their respective sentences, and apply self-attention to these tokens across all sentences. This allows the encoder to obtain the context information, and produce a set of contextually enriched `[CLS]` tokens,

[2] nltk.org

e_i^{CLS}, $1 \le i \le M$, each representing its sentence in the context of the whole document. Positional information is retained by adding the standard learnable position embeddings. The output from this transformer is h_i^{CLS}.

After capturing the context information, the second transformer encoder further refines each sentence's representation. It takes the context-enriched h_i^{CLS} token from the first encoder and combines it with the original embeddings of the sentence tokens. The combined input $[h_i^{CLS}, e_i^1, ..., e_i^L]$ undergoes another round of self-attention, producing the enriched sentence representation $[r_i^{CLS}, r_i^1, ..., r_i^L]$ where each r_i is influenced both by its original context and the broader document context.

3.3 Aspect Aggregation Block

This block plays a pivotal role in the DART framework by generating a unified document representation that captures the overall sentiment of a document concerning a specific aspect. It serves as a bridge between understanding individual sentences and comprehending the document as a whole, especially concerning a specific aspect. Given that sentiment towards an aspect can be scattered throughout a document, this block ensures that all these sentiments are appropriately aggregated.

The key idea of this block is obtain a aspect-specific representation through a two-level aggregation process. The first level weighs the importance of each token in the sentence concerning the aspect and the broader context by performing a local attentive pooling. The enriched sentence representation $[r_i^{CLS}, r_i^1, ..., r_i^L]$ from the global context interaction block undergoes a local aggregation process to obtain the output l_i:

$$l_i = \alpha_0 r_i^{CLS} + \sum_{k=1}^{L} \alpha_k r_i^k \qquad (1)$$

where α_k is the attention weight for the k^{th} token, determined based on its relevance to the aspect and the overall sentence context.

The second level takes the aggregated representations l_i of each sentence and performs a global attentive pooling to determine how much attention each sentence should receive when forming the overall document representation $\hat{d}_j$ with respect to the aspect a_j. This aggregation is given by

$$\hat{d}_j = \sum_{i=1}^{M} \frac{\exp(e_i^1 f(l_i))}{\sum_{i'=1}^{M} \exp(e_i^1 f(l_{i'}))} \ l_i \qquad (2)$$

where $f(\cdot)$ is a linear projection followed by the `tanh` function.

The weighting coefficients ensure that sentences more relevant to the aspect have a greater influence on the final document representation $\hat{d}_j$.

3.4 Sentiment Classification Block

This block is the final stage in the DART framework. The goal of this block is to utilize the aggregated document representation, which has been enriched with context and focused on a particular aspect, to predict the sentiment associated with that aspect. The final document representation $\hat{d}_j$ is passed through the two-layer MLP to obtain the probability distribution for the positive or negative sentiment towards the aspect a_j.

4 Performance Study

We implement DART in PyTorch1.13.0 and carry out experiments on the A100-SXM4 GPUs with 40 GB. We use the following datasets:

BeerAdvocate. This dataset contains reviews and ratings on predefined beer aspects: feel, look, smell, and taste, each rated on a scale of 1 to 5. The ratings are binarized into positive and negative sentiment.

TripAdvisor. This dataset consists of hotel reviews with ratings of 1 to 5 stars for aspects value, room, location, cleanliness, check in/front desk, service, and business. Again, these ratings are binarized.

SocialNews. We curate this dataset from news articles related to social issues from the PerSenT dataset (Bastan et al., 2020). We identify six implicit aspects, namely crime-justice, digital-online, economic issues, health, human rights, and work. A group of labelling experts was trained using educational guideline pack and a series of face to face sessions so that they have a clear understanding of the definition of aspects and sentiment. An expert benchmarking assessment was performed where 100 verified labels were assigned to each prospective annotators and those who reached a 70% agreement with experts were selected. Finally, three annotators are asked to assess the sentiment towards these aspects and we use the majority vote as the ground truth sentiment. The Kappa inter-annotator agreement is 93.14%.

Table 1 summarizes the characteristics of these datasets. DART utilizes the pre-trained model `bigbird-roberta-base` (Zaheer et al., 2020) in the Sentence Encoding Block. For the Global Context Interaction Block, the first Transformer en-

Dataset	#aspects	#docs	#long docs (%)	#sentences/doc	#tokens/doc	#tokens/sentence
BeerAdvocate	4	27583	217 (0.8%)	11.1	173.5	15.7
TripAdvisor	7	28543	4027 (14.1%)	12.9	298.9	23.1
SocialNews	6	4512	1031 (22.9%)	17.5	389.8	22.2

Table 1: Dataset characteristics. #long docs refers to documents with more than 512 tokens.

coder has 4 layers, while the second Transformer encoder has 2 layers. Both have 12 self-attention heads with a hidden size of 768. We use AdamW optimizer with a dropout rate of 0.1, and a batch size of 16. Each experiment is repeated 5 times and we report the average results on three datasets.

4.1 Comparative Study

We first compare DART with non-transformer aspect-based sentiment classification methods:

LRR (Wang et al., 2010) is a probabilistic graphical regression model. Guided by the overall rating and the aspect keywords, LRR infers the latent ratings for each aspect. A high rating indicates positive sentiment towards the aspect in the document.

VWS-DMSC (Zeng et al., 2019) is a weakly supervised model that predicts the sentiment with respect to an aspect. Target-opinion word pairs are extracted as supervision signal to learn the sentiment without using aspect polarity annotations.

D-MILN (Ji et al., 2020) is also a weakly supervised model for document-level aspect sentiment classification. It employs multiple instance learning to learn the relation between aspect-level and document-level sentiment.

N-DMSC (Yin et al., 2017) is a supervised neural model for document aspect sentiment classification. It employs hierarchical LSTM to generate aspect-aware document representations.

Table 2 shows the average accuracy for BeerAdvocate and TripAdvisor. We see that DART outperforms all the methods by a large margin. Using deep embedding features yields better results compared to traditional ngram features in LRR. Unlike N-DMSC, VWS-DMSC and D-MILN, DART does not require a pre-defined set of aspect-related keywords, reducing the complexity and pre-processing requirement in real-world scenarios.

Next, we compare the performance of DART with transformer-based models on long documents:

InstructABSA (Scaria et al., 2023) uses the 11B-parameter T5 model, with a maximum input sequence length of 512, for sentence-level aspect-based sentiment analysis. As such, we truncate the input when the length of instruction prompts and document exceeds 512 tokens.

Model	BeerAdvocate	TripAdvisor
LRR†	59.41	69.47
VWS-DMSC†	75.38	75.61
D-MILN†	79.86	79.52
N-DMSC†	86.35	83.34
DART	**88.25**	**86.38**

Table 2: Comparison of accuracy results for non-transformer models. Results with "†" are retrieved from (Ji et al., 2020).

MDABSA (Van Thin et al., 2022) is a joint multi-task architecture that aims to perform both aspect category detection and sentiment polarity classification tasks simultaneously.

Longformer (Beltagy et al., 2020) employs sliding windows to enable long-range coverage for long document modelling. We adapt Longformer for the document aspect sentiment classification task by first obtaining an aspect-aware document representation through feeding the aspect and document pair separated by the `[SEP]` token. The representation is then fed to a two-layer multi-layer perceptron to make sentiment prediction.

Big Bird (Zaheer et al., 2020) This is an encoder-only model that extends the sparse attention pattern with random attention for longer sequences. We adapt Bird Bird for sentiment classification in the same way as we have done for Longformer.

LongT5 (Guo et al., 2022) is the state-of-the-art transformer architecture for long inputs. The original LongT5 is an encoder-decoder structure with a new transient attention mechanism (TGlobal), which mimics ETC's local/global mechanism(Ainslie et al., 2020). Here, we leverage its encoder pre-trained weights and adapt it in the same way for fair comparison.

GPT4 (OpenAI, 2023) large language models (LLMs) have shown impressive results across various tasks. Here, we select `gpt-4-0613` as the representative LLM and perform the experiments under zero-shot and few-shot settings. For GPT4-

Model	All Aspects	Crime-Justice	Digital-Online	Economic Issues	Health	Human Rights	Work
InstructABSA	80.16	81.65	72.73	**81.25**	86.67	72.46	84.47
MDABSA	80.97	86.24	68.83	75.00	86.67	75.36	86.33
Longformer	80.53	87.89	69.09	78.75	80.67	70.72	85.71
Big Bird	80.81	86.97	69.35	76.25	79.33	75.36	86.09
LongT5	81.13	88.14	70.65	75.83	79.33	76.52	85.09
GPT4-zeroshot	58.91	72.48	25.97	58.33	66.67	63.77	62.11
GPT4-fewshot	60.32	75.23	28.57	64.58	70.00	65.22	60.25
DART	**83.81***	**88.53**	**75.64***	79.69	**89.17***	**78.99**	**86.80**

* indicates result is statistically significant when compared to the second best with p-value < 0.05.

Table 3: Accuracy of Transformer-based models in SocialNews Dataset.

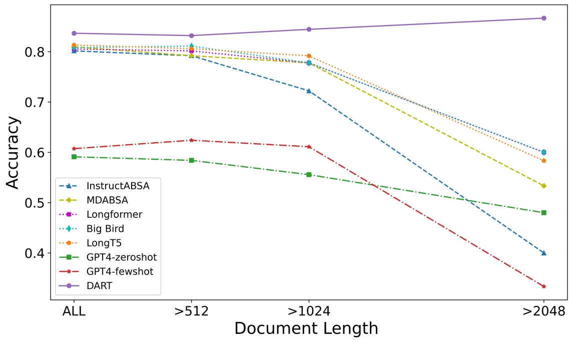


Figure 3: Accuracy of Transformer-based models with respect to document length on SocialNews test set.

Model	BeerAdvocate		TripAdvisor	
	Accuracy	F1	Accuracy	F1
InstructABSA	81.25	79.83	70.01	69.65
MDABSA	87.39	85.67	83.77	83.69
Longformer	87.85	86.09	83.61	83.13
Big Bird	88.14	86.59	84.04	83.44
LongT5	90.42	88.31	84.34	84.19
GPT4-zeroshot	58.54	51.09	59.39	57.65
GPT4-fewshot	69.65	66.71	74.43	72.16
DART	**94.44***	**92.86***	**86.48**	**85.96**

* indicates result is statistically significant when compared to the second best with p-value < 0.05.

Table 4: Comparison of results for transformer-based models on long documents (>512 tokens).

fewshot, we adopt the prompt from (Scaria et al., 2023) to perform aspect sentiment predictions.

Table 3 shows the performance of Transformer-based models on the SocialNews dataset, with details of their accuracy in handling different aspects. The results indicate that DART excels in five key aspects, particularly in the digital-online and health aspects, and is the second best model for the economic issues aspect. This demonstrates DART's ability to handle diverse and complex aspects. Appendix A provides a visualization of the learned document representations via t-sne.

Figure 3 shows the accuracy achieved in SocialNews for documents that exceed a certain length, as specified on the x-axis. The gap in performance between DART and other models widens as the document length surpasses the 1024-token threshold. DART continues to demonstrate superior performance even with extremely long documents, exceeding 2048 tokens in length. This indicates DART's proficiency in analyzing larger documents, which is an important aspect in real-world sentiment analysis scenarios.

Table 4 shows the average accuracy and macro F1 scores on the long documents in Beer Advocate and TripAdvisor over 5 runs. We see that DART achieves the best performance, with marked improvements over existing models. Similar gains is observed for the F1 scores, confirming DART's effectiveness in dealing with long documents for sentiment classification.

InstructABSA, which achieved state-of-the-art on SemEval 2014, 15, and 16 datasets for aspect sentiment classification, and MDABSA both perform worse than DART. This indicates that the methodologies developed for sentence-level aspect-based sentiment analysis or short texts do not extend well to longer documents. The results also reveal that our more compact, specialized DART model, which contains 687 million parameters, exceeded the performance of GPT4.

4.2 Ablation Study

We examine the effect of the various components in DART on its performance. We implemented two variants: (a) w/o Int. where the interaction component is bypassed and the outputs from the Sentence Encoding Block is fed directly to the Aspect Aggregation Block; and (b) w/o Agg. where the aggregation component is omitted and the average of the [CLS] vectors is used as the document representation for sentiment prediction.

Model	BeerAdvocate	TripAdvisor	SocialNews
w/o Int.	86.57	85.41	80.37
w/o Agg.	86.74	85.51	80.78
DART	87.94	86.21	85.54

Table 5: Accuracy of DART and its variants.

Table 5 shows the results. Compared to BeerAdvocate and TripAdvisor, we see a significant drop in the accuracy for SocialNews when the Sentence Interaction block is removed because 22.9% of the documents are longer than 512 on SocialNews. Similar reduction in accuracy is observed when we do not incorporate the Aspect Aggregation block. This demonstrates the importance of capturing the interaction among sentences in long documents as well as aggregating aspects locally and globally.

4.3 Case Studies

Here, we present case studies to show DART's ability to highlight phrases relevant to the target aspects. Figure 4 shows an article from SocialNews related to the aspect HEALTH. Only DART correctly predicts the negative sentiment towards this aspect while both Big Bird and Longformer give a positive sentiment. Phrases in purple are highlighted by DART as the basis for its negative prediction. In contrast, Big Bird and Longformer could not adequately capture the context, leading them to overlook the underlying negative sentiment.

Figure 5 shows two sample reviews from TripAdvisor. For the top review, DART focuses on phrases related to VALUE (highlighted in red) and correctly predicts a positive sentiment towards the aspect VALUE while Big Bird and Longformer give the wrong predictions. For the bottom review, DART predicts the correct negative sentiment towards the aspect CLEAN, with relevant phrases highlighted in green. We see that although DART attends to the phrase "Overall room be clean daily", it is able to identify negative phrases such as "exotic huge dead cockroach", "dingy bed and blanket" and "The shower stall do not close" to be associated to the aspect CLEAN and gives the correct prediction. In contrast, Big Bird and Longformer mistakenly interpret the sentiment as positive.

5 Application of DART to Trust and Polarity Prediction

While DART is originally conceptualized for sentiment analysis, the framework is versatile and can be extended for trust analysis and polarity prediction. In this section, we show that DART's ability to capture context information and aspect-specific attention aggregation makes it well-suited to evaluate trust-related aspects and gauge the degree of alignment or opposition on a topic.

Model	TrustData	Hyperpartisan
Longformer	80.77	93.54
Big Bird	81.59	92.00
LongT5	82.26	93.23
GPT4-zeroshot	77.89	83.08
GPT4-fewshot	79.95	86.15
DART	**83.93***	**95.69***

* statistically significant compared to the second best with p-value < 0.05.

Table 6: Accuracy of trust and polarity predictions.

Model	Ability	Dependability	Integrity	Purpose
Longformer	80.24	75.56	88.29	83.33
Big Bird	80.95	78.89	84.87	85.71
LongT5	81.43	78.89	88.78	85.23
GPT4-zeroshot	75.79	81.48	80.49	83.33
GPT4-fewshot	78.17	**85.19***	82.93	80.95
DART	**83.23***	81.48	**89.63**	**87.14**

* statistically significant compared to the second best with p-value < 0.05.

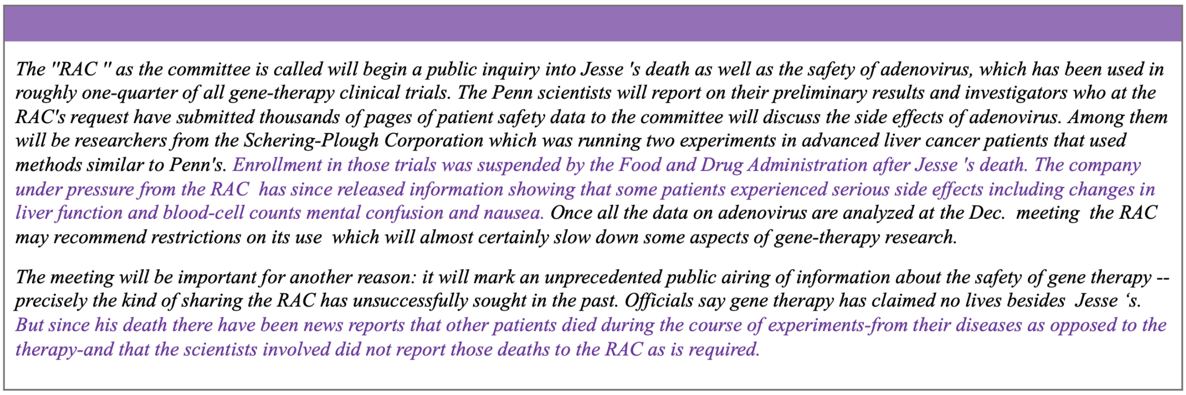
Table 7: Accuracy for various aspects in TrustData.

We compile a dataset for trust prediction, comprising of 2925 documents, of which 60.5% are long documents with more than 512 tokens. This dataset emphasizes four trust-related aspects: ability, dependability, integrity and purpose. We enlist three independent annotators to assess the trust polarity for each aspect, and take the majority vote as the ground truth. The annotation labels are "trust", "distrust", "mixed", and "no indication". The Kappa inter-annotator agreement is 87.29%. We call this dataset **TrustData**.

For polarity prediction, we use the **Hyperpartisan** dataset (Kiesel et al., 2019) consisting of news articles which have been manually labelled as hyperpartisan or not. There are 645 articles, out of which 53.3% have more than 512 tokens.

In Table 6, we see that DART gives the best accuracy and F1 score for trust and polarity predictions. The improvements achieved by DART over the second best model are statistically significant with p-value < 0.05, indicating the effectiveness of the global context interaction block in DART to capture the context information in long documents.

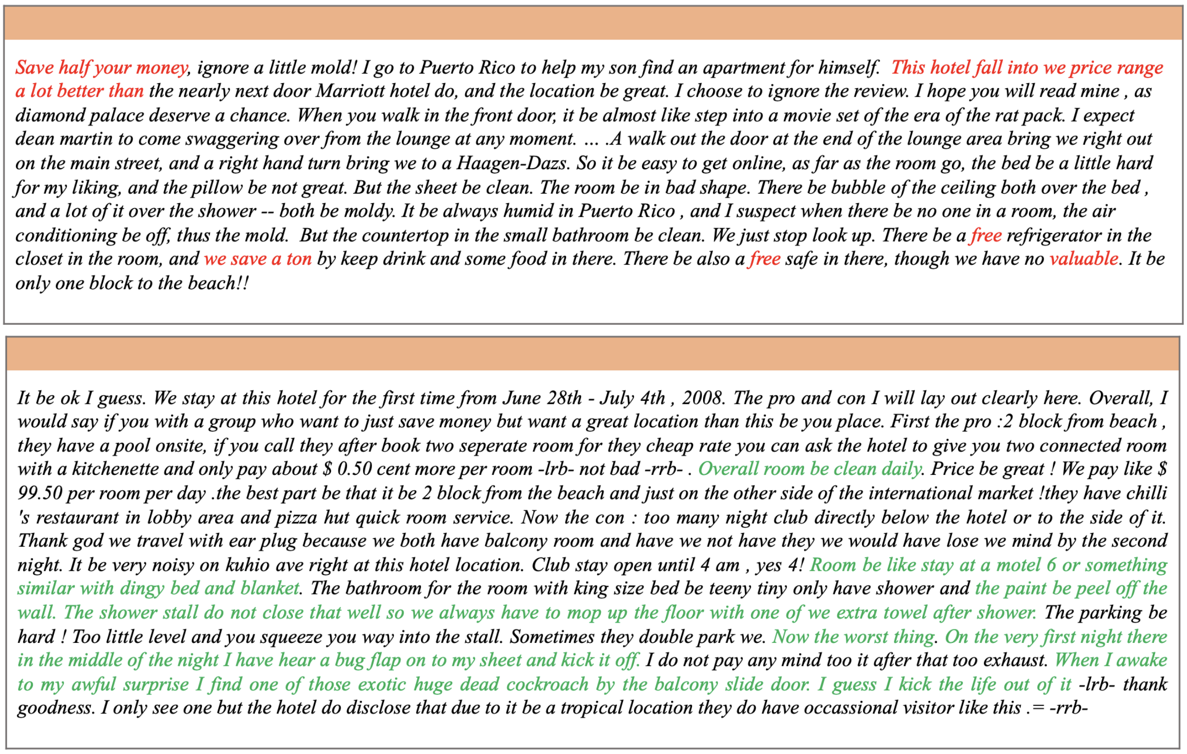
Table 7 provides a detailed breakdown of model accuracies in predicting the polarity of the different aspects in the TrustData dataset. We see that DART gives the best performance in three key aspects, and



The "RAC " as the committee is called will begin a public inquiry into Jesse 's death as well as the safety of adenovirus, which has been used in roughly one-quarter of all gene-therapy clinical trials. The Penn scientists will report on their preliminary results and investigators who at the RAC's request have submitted thousands of pages of patient safety data to the committee will discuss the side effects of adenovirus. Among them will be researchers from the Schering-Plough Corporation which was running two experiments in advanced liver cancer patients that used methods similar to Penn's. Enrollment in those trials was suspended by the Food and Drug Administration after Jesse 's death. The company under pressure from the RAC has since released information showing that some patients experienced serious side effects including changes in liver function and blood-cell counts mental confusion and nausea. Once all the data on adenovirus are analyzed at the Dec. meeting the RAC may recommend restrictions on its use which will almost certainly slow down some aspects of gene-therapy research.

The meeting will be important for another reason: it will mark an unprecedented public airing of information about the safety of gene therapy -- precisely the kind of sharing the RAC has unsuccessfully sought in the past. Officials say gene therapy has claimed no lives besides Jesse 's. But since his death there have been news reports that other patients died during the course of experiments-from their diseases as opposed to the therapy-and that the scientists involved did not report those deaths to the RAC as is required.

Figure 4: Sample article from SocialNews. DART correctly predicts positive sentiment for the HEALTH aspect while Big Bird and Longformer predicts negative sentiment.



Save half your money, ignore a little mold! I go to Puerto Rico to help my son find an apartment for himself. This hotel fall into we price range a lot better than the nearly next door Marriott hotel do, and the location be great. I choose to ignore the review. I hope you will read mine , as diamond palace deserve a chance. When you walk in the front door, it be almost like step into a movie set of the era of the rat pack. I expect dean martin to come swaggering over from the lounge at any moment. ... .A walk out the door at the end of the lounge area bring we right out on the main street, and a right hand turn bring we to a Haagen-Dazs. So it be easy to get online, as far as the room go, the bed be a little hard for my liking, and the pillow be not great. But the sheet be clean. The room be in bad shape. There be bubble of the ceiling both over the bed , and a lot of it over the shower -- both be moldy. It be always humid in Puerto Rico , and I suspect when there be no one in a room, the air conditioning be off, thus the mold. But the countertop in the small bathroom be clean. We just stop look up. There be a free refrigerator in the closet in the room, and we save a ton by keep drink and some food in there. There be also a free safe in there, though we have no valuable. It be only one block to the beach!!

It be ok I guess. We stay at this hotel for the first time from June 28th - July 4th , 2008. The pro and con I will lay out clearly here. Overall, I would say if you with a group who want to just save money but want a great location than this be you place. First the pro :2 block from beach , they have a pool onsite, if you call they after book two seperate room for they cheap rate you can ask the hotel to give you two connected room with a kitchenette and only pay about $ 0.50 cent more per room -lrb- not bad -rrb- . Overall room be clean daily. Price be great ! We pay like $ 99.50 per room per day .the best part be that it be 2 block from the beach and just on the other side of the international market !they have chilli 's restaurant in lobby area and pizza hut quick room service. Now the con : too many night club directly below the hotel or to the side of it. Thank god we travel with ear plug because we both have balcony room and have we not have they we would have lose we mind by the second night. It be very noisy on kuhio ave right at this hotel location. Club stay open until 4 am , yes 4! Room be like stay at a motel 6 or something similar with dingy bed and blanket. The bathroom for the room with king size bed be teeny tiny only have shower and the paint be peel off the wall. The shower stall do not close that well so we always have to mop up the floor with one of we extra towel after shower. The parking be hard ! Too little level and you squeeze you way into the stall. Sometimes they double park we. Now the worst thing. On the very first night there in the middle of the night I have hear a bug flap on to my sheet and kick it off. I do not pay any mind too it after that too exhaust. When I awake to my awful surprise I find one of those exotic huge dead cockroach by the balcony slide door. I guess I kick the life out of it -lrb- thank goodness. I only see one but the hotel do disclose that due to it be a tropical location they do have occassional visitor like this .= -rrb-

Figure 5: Sample reviews from TripAdvisor. Top: DART gives correct positive sentiment prediction for the aspect VALUE. Bottom: DART gives correct negative sentiment prediction for the aspect CLEAN.

is competitive with GPT4-zeroshot for the Dependability aspect, possibly due to the fewer number of documents with this aspect. The improvements suggest that the aspect-specific attention aggregation block in DART significantly enhances its ability to focus on phrases relevant to the various aspects.

6 Conclusion

We have described DART, a hierarchical transformer-based framework for document-level aspect-based sentiment analysis. DART handles the complexities of longer text through its global context interaction and two-level aspect aggregation blocks, which enhance the model's ability to recognize and amplify aspect-specific content across long-range dependencies. This enables DART to focus on relevant phrases associated with the target aspect. Experiments on various datasets indicate DART's effectiveness in handling long documents. We have also shown the applicability of DART for trust and polarity prediction and will make the curated SocialNews dataset publicly available. Future work includes extending DART's capabilities to handle aspect-based sentiment analysis involving multiple entities.

Acknowledgments

This work is supported by the Singapore Ministry of Education Academic Research Fund Tier 1. We would like to thank Katie Stuart for her comments.

Limitations

This work assumes each document contains only one entity. There is a need to develop a benchmark that can assess aspect based sentiments towards different entities in long documents.

References

Joshua Ainslie, Santiago Ontanon, Chris Alberti, Vaclav Cvicek, Zachary Fisher, Philip Pham, Anirudh Ravula, Sumit Sanghai, Qifan Wang, and Li Yang. 2020. ETC: Encoding long and structured inputs in transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 268–284, Online. Association for Computational Linguistics.

Mohaddeseh Bastan, Mahnaz Koupaee, Youngseo Son, Richard Sicoli, and Niranjan Balasubramanian. 2020. Author's sentiment prediction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 604–615, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

Gianni Brauwers and Flavius Frasincar. 2021. A survey on aspect-based sentiment classification. *ACM Computing Surveys (CSUR)*.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Recurrent neural network-based sentence encoder with gated attention for natural language inference. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pages 36–40, Copenhagen, Denmark. Association for Computational Linguistics.

Xiao Chen, Changlong Sun, Jingjing Wang, Shoushan Li, Luo Si, Min Zhang, and Guodong Zhou. 2020. Aspect sentiment classification with document-level sentiment preference modeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3667–3677, Online. Association for Computational Linguistics.

Shizhe Diao, Sedrick Scott Keh, Liangming Pan, Zhiliang Tian, Yan Song, and Tong Zhang. 2023. Hashtag-guided low-resource tweet classification. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 1415–1426, New York, NY, USA. Association for Computing Machinery.

Hao Fei, Jingye Li, Yafeng Ren, Meishan Zhang, and Donghong Ji. 2022. Making decision like human: Joint aspect category sentiment analysis and rating prediction with fine-to-coarse reasoning. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 3042–3051, New York, NY, USA. Association for Computing Machinery.

Hao Fei, Yafeng Ren, Shengqiong Wu, Bobo Li, and Donghong Ji. 2021. Latent target-opinion as prior for document-level sentiment classification: A variational approach from fine-grained perspective. In *Proceedings of the Web Conference 2021*, WWW '21, page 553–564, New York, NY, USA. Association for Computing Machinery.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. LongT5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736, Seattle, United States. Association for Computational Linguistics.

Sk Mainul Islam and Sourangshu Bhattacharya. 2022. AR-BERT: Aspect-relation enhanced aspect-level sentiment classification with multi-modal explanations. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 987–998, New York, NY, USA. Association for Computing Machinery.

Yunjie Ji, Hao Liu, Bolei He, Xinyan Xiao, Hua Wu, and Yanhua Yu. 2020. Diversified multiple instance learning for document-level multi-aspect sentiment classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7012–7023, Online. Association for Computational Linguistics.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Bin Lu, Myle Ott, Claire Cardie, and Benjamin K Tsou. 2011. Multi-aspect sentiment analysis with topic models. In *2011 IEEE 11th international conference on data mining workshops*, pages 81–88. IEEE.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4068–4074.

Julian McAuley, Jure Leskovec, and Dan Jurafsky. 2012. Learning attitudes and attributes from multi-aspect reviews. In *2012 IEEE 12th International Conference on Data Mining*, pages 1020–1025. IEEE.

Ambreen Nazir, Yuan Rao, Lianwei Wu, and Ling Sun. 2020. Issues and challenges of aspect-based sentiment analysis: a comprehensive survey. *IEEE Transactions on Affective Computing*.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8600–8607.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Kevin Scaria, Himanshu Gupta, Saurabh Arjun Sawant, Swaroop Mishra, and Chitta Baral. 2023. InstructABSA: Instruction learning for aspect based sentiment analysis. *CoRR*, abs/2302.08624.

Aliaksei Severyn and Alessandro Moschitti. 2015. Twitter sentiment analysis with deep convolutional neural networks. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 959–962.

Duyu Tang, Bing Qin, and Ting Liu. 2016. Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 214–224, Austin, Texas. Association for Computational Linguistics.

Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. 2018. Learning to attend via word-aspect associative fusion for aspect-based sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Peter Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 417–424, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Dang Van Thin, Lac Si Le, Hao Minh Nguyen, and Ngan Luu-Thuy Nguyen. 2022. A joint multi-task architecture for document-level aspect-based sentiment analysis in vietnamese. *IJMLC*, 12(4).

Hongning Wang, Yue Lu, and Chengxiang Zhai. 2010. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 783–792.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615, Austin, Texas. Association for Computational Linguistics.

Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2021. Hi-Transformer: Hierarchical interactive transformer for efficient and effective long document modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 848–853, Online. Association for Computational Linguistics.

Wei Xue and Tao Li. 2018. Aspect based sentiment analysis with gated convolutional networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2514–2523, Melbourne, Australia. Association for Computational Linguistics.

Hang Yan, Junqi Dai, Tuo Ji, Xipeng Qiu, and Zheng Zhang. 2021. A unified generative framework for aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2416–2429, Online. Association for Computational Linguistics.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

Yichun Yin, Yangqiu Song, and Ming Zhang. 2017. Document-level multi-aspect sentiment classification as machine comprehension. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2044–2054, Copenhagen, Denmark. Association for Computational Linguistics.

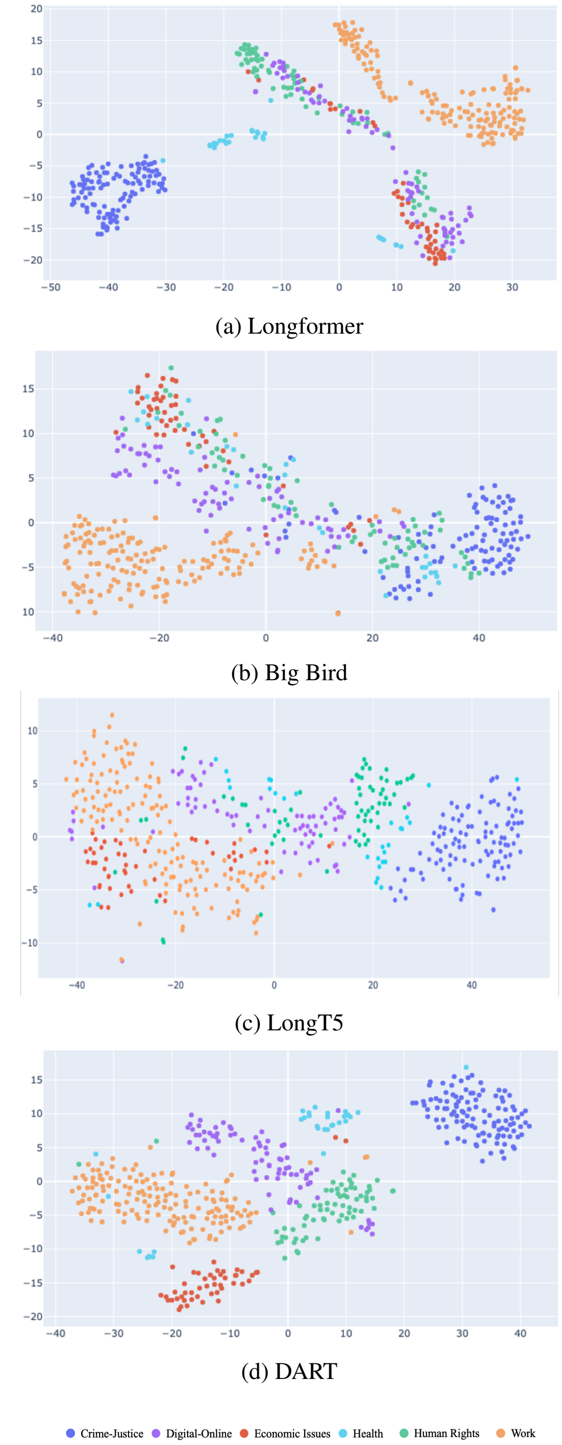
Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big Bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Ziqian Zeng, Wenxuan Zhou, Xin Liu, and Yangqiu Song. 2019. A variational approach to weakly supervised document-level multi-aspect sentiment classification. In *Proceedings of the 2019 Conference of* *the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 386–396, Minneapolis, Minnesota. Association for Computational Linguistics.

Meishan Zhang, Yue Zhang, and Duy-Tin Vo. 2016. Gated neural networks for targeted sentiment analysis. In *Thirtieth AAAI conference on artificial intelligence*.

A Visualization of Learned Representations

Figure 6 gives a visualization of the learned document representations via t-sne.



(a) Longformer

(b) Big Bird

(c) LongT5

(d) DART

Figure 6: t-SNE visualization of document representations for SocialNews.

We see that that DART's learned representation is well separated for the aspects digital-online, economic issues, and work occupation in the SocialNews dataset. In contrast, the learned representations of Longformer, Big Bird and LongT5 tend to be mixed and cannot distinguish between the different aspects.