

---

# Using Relational and Causality Context for Tasks with Specialized Vocabularies that are Challenging for LLMs

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

1 Short text is typical for reports such as incident synopsis and product feedback  
2 for efficiency and convenience. However, classifying short reports can be very  
3 challenging due to incomplete information and limited labeled data, and in some  
4 cases, many domain-specific terms. To address these issues, we examine the use of  
5 causality, as represented by linguistic cause and effect, in models for short report  
6 classification. We propose two augmentations of a hierarchical graph attention  
7 network to represent latent causes and effects. We also investigate the effectiveness  
8 of using a pretrained Language Model SBERT vs. the more traditional tf-idf repre-  
9 sentations for reports with general and specialized vocabularies. Experiments on  
10 five public report datasets verify that inclusion of causality in modeling short report  
11 datasets with many domain-specific terms improves classification performance.

## 12 1 Introduction

13 Reporting delivers important information for many critical circumstances and is essential to inform  
14 a variety of decision making. For example, regulators use reports to analyze causes of problems  
15 in accidents to determine responsibilities and prevention strategies [8]. Another example includes  
16 business reports to review feedback of products to understand satisfaction and unmet needs [12].  
17 Categorizing short text reports can take tremendous manual effort and expert knowledge, hence it  
18 is frequently impractical to sort and label every report. This raises a strong need for the research of  
19 short text report classification.

20 The challenges of short text classification include lacking sufficient labels for training a classifier [5]  
21 and lacking context in short text content [19]. Large Language Models (LLMs) such as Sentence  
22 Transformer (SBERT) [20] has been used for few-shot learning to adapt to new tasks with label  
23 scarcity. Because these LLMs are pretrained by a large amount of web data, they maintain a strong  
24 understanding of context for a wide array of short text classification. However, in some tasks like  
25 short report classification, the report text can include many domain-specific terms that are not well  
26 represented in pretrained LLM models.

27 Another thread of research uses graph neural networks to capture relationships and context in  
28 relational data for learning rich and task-specific representation. Prior work [16] leverages latent  
29 topics and entity recognized in text content to enrich the context in heterogeneous graphical attention  
30 network (HGAT) for semi-supervised learning. However, while the general latent topics are relevant,  
31 they may not be precise enough context to classify reports. For example, in a task of classifying  
32 traffic incident reports, a popular word “car” can be relevant to many incidents but is not precise to  
33 indicate problems or consequences of a traffic incident.

34 Fortunately, text written in many short reports tend to include language with causal relations. For  
35 example, railroad incident reports presented in Figure 1 include abnormal weather patterns (e.g.,  
36 “extreme environmental conditions wind velocity”) or operation failures (e.g., “failed to be in proper  
37 position”) that lead to certain situations (e.g., “going in the wrong direction,” “derailed 9 cars”). This  
38 explicit or implicit causality expressed in text content may help classify reports into categories that  
39 are important for investigation, e.g., accident causes or side effects. We hypothesize that the text that  
40 indicates likely cause and effect can provide more precise context for report analysis.

41 In this work, we propose to augment graph neural networks with causality context for semi-supervised  
42 classification on short text reports. This approach allows information propagation to be more aligned  
43 with potential causes and effects indicated in text content. Our contributions include: (1) We  
44 investigate the effectiveness of using an LLM, SBERT vs. the traditional tf-idf representations for  
45 classifying short reports with general and specialized vocabularies. (2) We propose a framework that  
46 captures the context and aligns connections in a graph neural network with relevant causality relations  
47 to improve short report classification. (3) We evaluate the proposed idea over the state-of-the-art  
48 across five public datasets, where it consistently enhances the classification targets with specialized  
49 vocabularies.

## 50 **2 Related Work**

51 A popular research question in text classification is the strategies to leverage labels for training. Pre-  
52 trained language models such as SBERT [20] and RoBERTa [17] harness self-attention mechanisms  
53 to capture contextual relationships from large scale web text. These pretrained LLMs have been  
54 demonstrated effective for few-shot text classification for the tasks with vocabularies well represented  
55 in web data (e.g., news and online reviews) that these pretrained models are trained on.

56 In addition, semi-supervised learning is effective for use cases with label scarcity [18]. Many  
57 approaches are proposed based on graph neural network models (GNNs) because GNNs can leverage  
58 the inherent structure of the graph data such as GCN [15] and TextGCN [27], which contains  
59 relationships between data points. Graph convolution networks (GCN) [15] is an efficient variant of  
60 convolutional neural networks which operate directly on data represented as graphs. The concept of  
61 GCN is applied for semi-supervised text classification (TextGCN) [27] by modeling the text corpus  
62 as a document-work graph. Heterogeneous attention networks [16, 25, 10] are designed to handle  
63 heterogeneous graphs where nodes belong to different types, and provides an attention mechanism to  
64 capture relevant context from different node types to address the challenge of missing context. Our  
65 work uses it as the backbone and distills the relational structure by infusing the context of causality  
66 inferred in text to improve report classification.

67 Causality is a metaphysical concept that is commonly seen in real world text data. A typical causal  
68 relation in text refers to a relationship between text arguments where one (cause) is responsible for  
69 causing the other (effect) [6]. Understanding causality has been explored by many natural language  
70 processing works [13, 11]. Particularly, large causal datasets [23, 22, 21] have been released for  
71 training and verification, which enabled much progress in cause-effect-signal span detection. Our  
72 work uses the detector to extract specific spans of text within a given report that represent the cause  
73 and effect. Since the extracted text can be noisy due to detection errors and diverse expressions, we  
74 use latent topics of the extracted cause and effect text to represent causality context and build the  
75 relational structure.

## 76 **3 Our Proposed Approach**

77 First, we introduce the approach to capture the representative context of causality, namely, latent  
78 causes and effects, from short text reports. Second, we present a framework to add latent causes and  
79 latent effects to a graph neural network to improve report classification.

### 80 **3.1 Representing Latent Causes and Effects**

81 Our goal is to provide representative cause and effect context in a graph neural network to model  
82 more precise relationships. To learn representative causes and effects, we detect the causal relation at

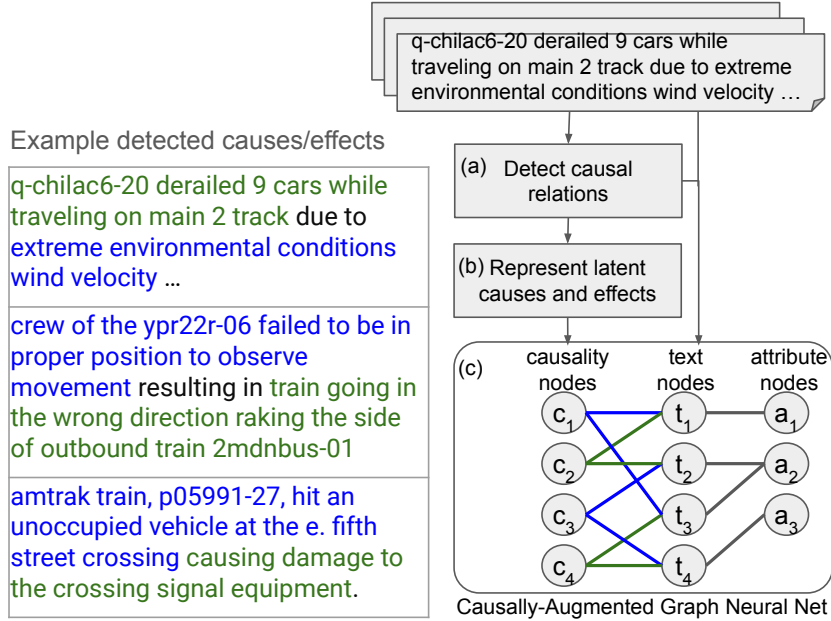


Figure 1: Left: Example cause text (blue) and effect text (green) detected from incident reports. Right: System diagram of our approach. Best seen in color.

83 the level of each text report (Figure 1 (a)), and then aggregate similar cause and effect text extracted  
 84 from different reports into representative latent causes and effects (Figure 1 (b)).

85 A causal relation presented in free text can be signaled in different ways; for example, “due to” in the  
 86 first report of Figure 1 and the alternatives such as “resulting in” and “causing” as appeared in the  
 87 second and third report, respectively. Prior work has attempted to collect annotated text and train  
 88 models to classify causal events or detect cause and effect signal span. Tan et al. [23, 22, 21] use  
 89 a neural network to train a cause-effect-signal span detector based on annotated news dataset. We  
 90 leverage the detector  $D$  to extract the text segments that are likely to indicate a cause or a effect,  
 91 denoted as  $(t_n^x, t_n^e) = D(t_n)$ , where  $t_n^x$  is the cause text segment and  $t_n^e$  is the effect text segment  
 92 extracted from a report  $t_n$  as presented in Figure 1.

93 The extracted cause and effect text are inherently noisy due to diverse expressions and detection  
 94 errors. For example, “a go-around event” and “a change of route” may be referred to as the same  
 95 type of event. We use LDA [7] to discover latent topics  $\theta$  from extracted cause text and effect text as  
 96 latent causes and effects. Each of latent causes and effects is represented by the word distribution  
 97  $(w_1, \dots, w_m)$ .  $m$  is the vocabulary size.

### 98 3.2 Causally-Augmented Graph Neural Net

99 We include the latent causes and effects in a graph neural network to align the information propagation  
 100 with similar causal relations in our causally-augmented model, **CHGAT**. Our graph  $G = (V, E)$   
 101 where  $V$  and  $E$  represent nodes and edges respectively has heterogeneous types of nodes, including  
 102 causality nodes ( $c_p$ ), text report nodes ( $t_n$ ), and attribute nodes ( $a_q$ ) as shown in Figure 1 (c). Causality  
 103 nodes include latent causes and effects, where a node  $c_p$  corresponds to  $\theta$  (cf. Section 3.1) and is  
 104 represented by its word distribution. We use topic distribution inference to calculate the posterior  
 105 probabilities  $p(\theta_{c_p} | t_n^x)$  and  $p(\theta_{c_p} | t_n^e)$  for each pair  $c_p$  and  $t_n$ . The edges between  $t_n$  and the  $k$   
 106 causality nodes of  $c$  with the highest probability are initialized with 1, otherwise by 0, in an adjacency  
 107 matrix  $A$ . Self-connections are also initialized as 1 in  $A$ . We use HGAT to learn the attention weights  
 108 of heterogeneous types of nodes, where the layer-wise propagation rule is formulated as,

$$H^{(l+1)} = \sigma(\tilde{A} \cdot H^{(l)} \cdot W^{(l)}), \quad (1)$$

Table 1: Statistics of evaluation datasets.

Dataset	ASR	REA	ECD	WCR	DRU
#docs	13,090	10,000	480	19,663	4,045
#classes	2	5	6	5	5
#tokens	34.73	44.54	11.24	6.29	43.85
	$\pm 16.2$	$\pm 32.8$	$\pm 1.7$	$\pm 2.4$	$\pm 46.3$
domain	aviation	railroad	workplace	e-commerce	drug
no c/e	0.26	0.12	0.00	0.002	0.16

109  $\tilde{A} = M^{-\frac{1}{2}}AM^{-\frac{1}{2}}$  is the symmetric normalized adjacency matrix, where  $M_{ii} = \sum_j A_{ij}$  is a degree  
 110 matrix.  $W^{(l)}$  is a trainable transformation matrix for each layer.  $\sigma(\cdot)$  is an activation function such as  
 111 ReLU. After an  $L$ -layer HGAT with the embeddings of all nodes, the embedding of text nodes are  
 112 fed to a softmax layer for classification.

113 For a fair comparison, the text embedding of  $t_n$  is initialized with tf-idf, as used in HGAT, but it can  
 114 be replaced with other text embeddings. HGAT uses name entity recognition to detect entities in text  
 115 to represent attribute nodes. Since name entities are less common in report data than news and social  
 116 media data that the prior work mainly addressed, we instead use report attributes (e.g. reporting  
 117 office) that are already associated with each report to represent the attribute nodes  $a_q$ , where each  
 118 attribute is initialized with a one-hot vector. The same attribute nodes are used in both the baseline  
 119 HGAT and the proposed approach in our experiments. The report attributes used in different datasets  
 120 are reported in Appendix A.

## 121 4 Experiments

122 The evaluation aims to investigate (1) the effectiveness of LLM vs. non-LLM representations for  
 123 classifying short text reports with general and specialized vocabulary, and (2) the effectiveness of  
 124 relational and causality context for short reports with vocabularies that are not well represented in  
 125 LLMs. We first present the experiment settings and then discuss the experimental results.

### 126 4.1 Datasets

127 The evaluation is conducted on the 5 public datasets below. We report data statistics in Table 1  
 128 including the proportion of text samples without cause and effect text detected (no c/e). Three  
 129 of these datasets (ASR, REA, DRU) include terms specific to a certain knowledge domain, i.e.,  
 130 aviation, railroad and drug. For each dataset, we randomly select 40 labeled samples per class,  
 131 20 of which for training and the rest for validation. Similar to [15], all the left samples are  
 132 used for testing and are also used as unlabeled samples for training. In our experiments, all  
 133 conditions are tested once. More details of the datasets (e.g., class labels) are presented in Appendix A.

134  
 135 **ASR:** This dataset [1] includes reports published in the Aviation Safety Reporting System  
 136 database from NASA. We use the column of situations to generate binary class labels, ie. related to  
 137 human factors or not. Data entries without clear cue to determine the labels are removed. This dataset  
 138 includes many terms specific to the aviation domain.

139 **REA:** This dataset [3] includes railway incident reports published by the Federal Railroad  
 140 Administration, Office of Railroad Safety. We use the top five accident cause codes as the class labels,  
 141 and randomly sample 2,000 reports per class for experiments. This dataset includes many terms  
 142 specific to the railroad traffic and incidents.

143 **ECD:** This dataset [2] includes employee complaints in an organization. We use top 5 complaint  
 144 genres as the class labels and one more class label to include any other complaint genres.

145 **WCR:** This is a Women’s Clothing E-Commerce dataset with customers’ reviews [4]. Since the full  
 146 review text is long (mean: 72.15 tokens), we use the review titles as text reports. We use the 5 levels  
 147 of satisfaction ratings as the class labels.

148 **DRU:** The dataset [14] includes patient reviews on specific drugs along with related side effects. We  
 149 use the 5 levels of side effects as the class labels. This dataset includes many specific drugs along  
 150 with related conditions.

Table 2: Test accuracy (A) and macro-F1 (F) of the baselines and our approach. The best results are bolded.

Dataset		SBERT	tf-idf			
		SVM	SVM	HGAT	CHGAT	CHGAT+
ASR	A	48.50	58.70	67.96	67.10	<b>69.82</b>
	F	36.50	54.65	67.49	65.56	<b>69.39</b>
REA	A	27.86	25.41	48.43	48.18	<b>51.24</b>
	F	24.06	23.46	48.40	48.26	<b>51.63</b>
ECD	A	<b>82.50</b>	60.83	61.67	63.75	64.17
	F	<b>77.18</b>	54.00	57.89	61.39	61.30
WCR	A	<b>40.87</b>	19.21	24.60	28.33	26.51
	F	<b>30.33</b>	15.19	19.32	22.85	20.76
DRU	A	31.50	32.20	29.10	31.42	<b>35.81</b>
	F	25.79	27.71	26.88	28.05	<b>30.62</b>

Table 3: Test accuracy (A) and macro-F1 (F) for our approach with the use of either cause or effect context.

Dataset		CHGAT		CHGAT+	
		cause	effect	cause	effect
ASR	A	61.58	67.02	66.46	<b>68.22</b>
	F	61.11	65.02	66.46	<b>67.69</b>
REA	A	<b>48.46</b>	44.16	47.70	47.27
	F	<b>48.18</b>	44.14	47.66	47.42
ECD	A	61.25	60.00	<b>65.42</b>	64.17
	F	57.82	58.48	60.79	<b>62.76</b>
WCR	A	27.47	<b>36.88</b>	27.72	27.97
	F	21.90	<b>24.33</b>	21.46	22.58
DRU	A	29.86	32.07	35.03	<b>35.55</b>
	F	28.16	30.30	30.66	<b>30.82</b>

## 151 4.2 Approaches

152 We compare the approaches using LLM and non-LLM representations in a few-shot learning  
 153 framework, a state-of-the-art graph neural network for heterogeneous types of nodes and our  
 154 proposed approach that leverages both relational and causality context.

155  
 156 **SVM+SBERT:** This approach uses a pretrained SBERT to represent text data. The SVM  
 157 [9] classifier is trained with the labeled data only, representing a baseline without considering the  
 158 unlabeled data in a specified task.

159 **SVM+tfidf:** This baseline is similar to SVM+SBERT but uses tf-idf as text representations.

160 **HGAT:** HGAT [16] has been compared with a variety of graph neural networks such as TextGCN  
 161 [27], HAN [25], GCN [15], GAT [24] in the prior work [26] and obtained the state-of-the-art results  
 162 for short text classification.

163 Our approaches, **CHGAT** and **CHGAT+**, both use HGAT as the backbone with representations of  
 164 causality context. CHGAT+ is similar to CHGAT but includes the whole text  $t_n$  in addition to cause  
 165 or effect text, attempting to remedy the missing context due to undetected or inaccurately detected  
 166 cause-effect text spans. The SVM classifiers in the evaluation uses SBERT [20] and tf-idf as the text  
 167 embedding. The tf-idf representation is used in the approaches that leverage relational and causality  
 168 context i.e. HGAT, CHGAT, and CHGAT+, to differentiate context from the relationships in the  
 169 task-specific data versus the context from the pretrained LLM model. All of the approaches including  
 170 SVM use both text and attribute information. Parameter settings are reported in Appendix C.

Text reports	Predicted classes
Employees' skill gaps hinder team performance.	Lack of training
Micromanagement is making the workload worse.	Workload & Stress

Figure 2: Example predicted results.

## 171 4.3 Experiment Results

### 172 4.3.1 Comparison over LLM based and non-LLM based Text Representations

173 Table 2 reports the accuracy and F1 (macro-F1) of different approaches across 5 datasets. SBERT  
 174 helps improve the performance of the SVM classifiers on ECD and WCR, having a characteristic that  
 175 includes more general vocabulary than others. On the other hand, the models using tf-idf features  
 176 perform better or competitively on ASR, REA and DRU with many domain-specific terms(e.g., the  
 177 term *YNP04R-25* in REA). This implies that it is better to select a text representation method which  
 178 takes care of domain specific terms directly when the dataset has a large vocabulary of them.

### 179 4.3.2 Effectiveness of Relational and Causality Context

180 Table 2 shows that HGAT improves SVM+tfidf in the ASR and REA datasets, demonstrated the  
 181 effectiveness of relational context learned in graph neural networks. However, HGAT does not  
 182 improve in the DRU dataset. On the other hand, our proposed approach CHGAT+ consistently  
 183 outperforms the baseline models SVM and HGAT in both measures for the datasets with specialized  
 184 vocabularies, which suggests the effectiveness of causality context inferred from the linguistic  
 185 characteristics and the graph structure.

186 In addition, CHGAT+ performs better than CHGAT in more of the datasets. Particularly, for the  
 187 datasets with higher proportion of undetected cause-effect text spans (cf. no c/e in Table 1) such  
 188 as ASR, REA and DRU, CHGAT+ obtains more improvements. This suggests that CHGAT+ may  
 189 be a potential remedy for the missing context from imperfect cause-effect text span detection, by  
 190 considering whole text in addition to cause and effect text.

191 While there are missing cause and effect text, all the datasets have a majority of data samples with  
 192 cause and effect text detected. The highest undetected rate is 0.26 as reported in Table 1, suggesting  
 193 that it is possible to get plentiful causality context in short text reports. Note that, the F1 of cause-effect  
 194 text span detector is reported in the prior work [23], around 60%-70%.

### 195 4.3.3 Discussion

196 The effectiveness of cause and effect context varies over tasks as reported in Table 3. Cause context is  
 197 most effective in REA, where the class labels are accident causes that have more obvious relationships  
 198 with cause text. Effect context is more useful in other classification tasks such as side effect and  
 199 satisfaction. The trend of effectiveness appears in CHGAT and CHGAT+ are roughly consistent.

200 There is no clear winner between cause only and effect only models. Hence, including both cause and  
 201 effect and using a mechanism to estimate the importance of each could be useful. This work leverages  
 202 the attention mechanism to determine the importance of individual links to different causality nodes.  
 203 The mechanism offers flexibility to incorporate cause or effect for different tasks or mixed situations,  
 204 e.g., different classes (Figure 2) can be more related to cause text (blue) or effect text (green).

## 205 5 Conclusion

206 Our investigation found that the short text reports with specialized vocabularies are more challenging  
 207 for LLM based text representation. We propose to infuse the causality context from free text to a  
 208 graph neural network for short report classification. The experiments on five public datasets suggest  
 209 that relational and causality context improves short report classification for the data with specialized  
 210 vocabularies. In the future we will improve the mechanism of representing and attending causality  
 211 context to generalize its use and to improve large language models for tasks with domain-specific  
 212 text.

## References

- 213
- 214 [1] Dataset card for asrs aviation incident reports. <https://huggingface.co/datasets/elihoole/asrs-aviation-reports>. Accessed: 2024-02-06.
- 215
- 216 [2] Employee complaints: Voicing concerns. <https://www.kaggle.com/datasets/omarsobhy14/employee-complaints>. Accessed: 2024-02-06.
- 217
- 218 [3] Railroad accident & incident data. <https://www.kaggle.com/datasets/chrico03/railroad-accident-and-incident-data>. Accessed: 2024-02-06.
- 219
- 220 [4] Women's e-commerce clothing reviews. <https://www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews>. Accessed: 2024-02-06.
- 221
- 222 [5] Charu C. Aggarwal and ChengXiang Zhai. *A Survey of Text Classification Algorithms*, pages 163–222. Springer US, Boston, MA, 2012.
- 223
- 224 [6] Biswanath Barik, Erwin Marsi, and Pinar Öztürk. Event causality extraction from natural science literature. *Res. Comput. Sci.*, 117:97–107, 2016.
- 225
- 226 [7] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.
- 227
- 228 [8] Stephan Corrie and Stephan Corrie. The u.s. aviation safety reporting system. *World Aviation Congress*, 1997.
- 229
- 230 [9] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.
- 231
- 232 [10] Yong Dai, Linjun Shou, Ming Gong, Xiaolin Xia, Zhao Kang, Zenglin Xu, and Daxin Jiang. Graph fusion network for text classification. *Knowledge-Based Systems*, 236:107659, 2022.
- 233
- 234 [11] Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158, 2022.
- 235
- 236
- 237
- 238
- 239 [12] Anders P. Fundin and Bo L.S. Bergman. Exploring the customer feedback process. *Measuring Business Excellence*, 7(2), 2003.
- 240
- 241 [13] Yohan Jo, Seojin Bang, Chris Reed, and Eduard Hovy. Classifying argumentative relations using logical mechanisms and argumentation schemes. *Transactions of the Association for Computational Linguistics*, 9:721–739, 2021.
- 242
- 243
- 244 [14] Surya Kallumadi and Felix Grer. Drug Review Dataset (Druglib.com). UCI Machine Learning Repository, 2018. DOI: <https://doi.org/10.24432/C55G6J>.
- 245
- 246 [15] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.
- 247
- 248 [16] Hu Linmei, Tianchi Yang, Chuan Shi, Houye Ji, and Xiaoli Li. Heterogeneous graph attention networks for semi-supervised short text classification. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4821–4830, Hong Kong, China, November 2019. Association for Computational Linguistics.
- 249
- 250
- 251
- 252
- 253
- 254 [17] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019.
- 255
- 256
- 257 [18] Yue Lu and Chengxiang Zhai. Opinion integration through semi-supervised topic modeling. In *Proceedings of the 17th International Conference on World Wide Web, WWW '08*, page 121–130. Association for Computing Machinery, 2008.
- 258
- 259

- 260 [19] Xuan-Hieu Phan, Le-Minh Nguyen, and Susumu Horiguchi. Learning to classify short and  
 261 sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the*  
 262 *17th International Conference on World Wide Web, WWW '08*, page 91–100. Association for  
 263 Computing Machinery, 2008.
- 264 [20] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-  
 265 networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language*  
 266 *Processing*. Association for Computational Linguistics, 11 2019.
- 267 [21] Fiona Anting Tan, Hansi Hettiarachchi, Ali Hürriyetoğlu, Tommaso Caselli, Onur Uca,  
 268 Farhana Ferdousi Liza, and Nelleke Oostdijk. Event causality identification with causal news  
 269 corpus - shared task 3, CASE 2022. In *Proceedings of the 5th Workshop on Challenges and Ap-*  
 270 *plications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 195–208,  
 271 Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational  
 272 Linguistics.
- 273 [22] Fiona Anting Tan, Hansi Hettiarachchi, Ali Hürriyetoğlu, Nelleke Oostdijk, Onur Uca, Surendra-  
 274 brikram Thapa, and Farhana Ferdousi Liza. Event causality identification - shared task 3,  
 275 CASE 2023. In Ali Hürriyetoğlu, Hristo Tanev, Vanni Zavarella, Reyhan Yeniterzi, Erdem  
 276 Yörük, and Milena Slavcheva, editors, *Proceedings of the 6th Workshop on Challenges and*  
 277 *Applications of Automated Extraction of Socio-political Events from Text*, pages 144–150, Varna,  
 278 Bulgaria, September 2023. INCOMA Ltd., Shoumen, Bulgaria.
- 279 [23] Fiona Anting Tan, Hansi Hettiarachchi, Ali Hürriyetoğlu, Nelleke Oostdijk, Tommaso Caselli,  
 280 Tadashi Nomoto, Onur Uca, Farhana Ferdousi Liza, and See-Kiong Ng. Recess: Resource  
 281 for extracting cause, effect, and signal spans. In *Proceedings of the 13th International Joint*  
 282 *Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter*  
 283 *of the Association for Computational Linguistics*, pages 66–82, Nusa Dua, Bali, November  
 284 2023. Association for Computational Linguistics.
- 285 [24] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua  
 286 Bengio. Graph attention networks. In *International Conference on Learning Representations*  
 287 *(ICLR)*, 2018.
- 288 [25] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. Het-  
 289 erogeneous graph attention network. In *The World Wide Web Conference*, page 2022–2032.  
 290 Association for Computing Machinery, 2019.
- 291 [26] Tianchi Yang, Linmei Hu, Chuan Shi, Houye Ji, Xiaoli Li, and Liqiang Nie. Hgat: Heteroge-  
 292 neous graph attention networks for semi-supervised short text classification. *ACM Trans. Inf.*  
 293 *Syst.*, 39(3), may 2021.
- 294 [27] Liang Yao, Chengsheng Mao, and Yuan Luo. Graph convolutional networks for text classi-  
 295 fication. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and*  
 296 *Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Sym-*  
 297 *posium on Educational Advances in Artificial Intelligence, AAAI'19/IAAI'19/EAAI'19*. AAAI  
 298 Press, 2019.

## 299 A Datasets

300 This section includes more details about the datasets, including the class labels and the report attributes  
 301 used for attribute nodes  $a_q$ .

302 **ASR:** This dataset [1] includes reports published in the Aviation Safety Reporting System database  
 303 maintained by NASA. We use the column of contributing factors / situations to generate class labels.  
 304 This column has many data entries assigned with multiple labels. Since we only focus on multi-class  
 305 classification rather than multi-label classification, we only keep the data entries that is labeled with  
 306 “human factors” without any other labels, and the data entries that are labeled with any labels except  
 307 “human factors”. This results in a 2 class labels, i.e., “human factors” or “not human factors”. In  
 308 addition to that, the attributes in the columns of aircraft flight phase and aircraft make model name  
 309 are used for the attribute nodes. The column of synopsis is used for the text nodes.



310 **REA:** This dataset [3] includes railway incident reports published by the Federal Railroad Administra-  
311 tion, Office of Railroad Safety; contains data on railway incidents from 1975 to 2022. We use the top  
312 five accident cause codes as the class labels, i.e., “wide gage,” “switch improperly lined,” “shoving  
313 movement,” “switch point worn or broken,” “buffing or slack action excessive.” The attributes in the  
314 columns of reporting railroad code and report year are used for the attribute nodes. The column of  
315 narrative is used for the text nodes.

316 **ECD:** This dataset [2] includes employee complaints in an organization. We use top 5 complaint  
317 genres as the class labels and one more class label to include any other complaint genres, totally 6  
318 classes including “communication issues,” “workload and stress,” “management lifestyle,” “lack of  
319 training and development,” “work place environment” and “others”. The attributes in the columns of  
320 employee role and gender are used for the attribute nodes. The column of report is used for the text  
321 nodes.

322 **WCR:** This is a Women’s Clothing E-Commerce dataset with the reviews written by customers [4].  
323 The data has been anonymized, and references to the company in the review text have been replaced  
324 with “retailer.” We use the 5 level of review ratings as the class labels. The attributes in the columns  
325 of product class (type) and department are used for attribute nodes. The column of title is used for  
326 the text nodes.

327 **DRU:** The dataset [14] includes patient reviews on specific drugs along with related conditions and  
328 side effects. The data was obtained by crawling online pharmaceutical review sites. We use the 5 level  
329 of side effects as the class labels ranging from “No Side Effects”, “Mild Side Effects,” “Moderate  
330 Side Effects,” “Severe Side Effects,” “Extremely Severe Side Effects.” The attributes in the columns  
331 of drug name and effectiveness are used for the attribute nodes. The column of side effect reviews is  
332 used for the text nodes.

333 For all the datasets, data entries with missing class labels, text reports for text nodes and attributes for  
334 attribute nodes are removed.

## 335 **B Input feature for SVM**

336 For fair comparison, SVM has used attribute information as suggested in the prior work [16]. We  
337 convert attribute information into one-hot vector, then the vector is concatenated with tf-idf (SVM+tf-  
338 idf) or SBERT (SVM+SBERT) features derived from text nodes.

## 339 **C Parameter Settings**

340 We set  $k = 2$  for initializing the number of edges between a text node and its top-k causality nodes  
341 with 1. We set the number of LDA topics as 15, the layer number as 2, and the hidden dimension as  
342 512, same with the setting for HGAT. The learning rate is 0.01, and the dropout rate is 0.95.

## 343 **D Packages Used**

344 We use NLTK for preprocessing. In addition, we use scikit-learn for extracting tf-idf features and  
345 training LDA and SVM models.