Classification-Aware Neural Topic Model Combined With Interpretable Analysis - For Conflict Classification

Tianyu Liang[1], Yida Mu [1], Soonho Kim[2], Darline Larissa Kengne Kuate[2], Julie Lang[2], Rob Vos[2], Xingyi Song[1]

[1]Department of Computer Science, University of Sheffield, Sheffield, United Kingdom
[2]International Food Policy Research Institute, Washington, DC, USA
{tliang9, y.mu, x.song}@sheffield.ac.uk

Abstract

A large number of conflict events are affecting the world all the time. In order to analyse such conflict events effectively, this paper presents a Classification-Aware Neural Topic Model (CANTM-IA) for Conflict Information Classification and Topic Discovery. The model provides a reliable interpretation of classification results and discovered topics by introducing interpretability analysis. At the same time, interpretation is introduced into the model architecture to improve the classification performance of the model and to allow interpretation to focus further on the details of the data. Finally, the model architecture is optimised to reduce the complexity of the model.

1 Introduction

Hundreds of conflicts break out every day around the world, many of which have a major impact on the world's political and economic situation. A recent example is Ukraine Crisis, which has caused energy scarcity in Europe, a reduction in world food production and many other repercussions. For governments and institutions such as the IFPRI, the impact of conflict events can be greatly reduced if they are classified, analysed and responded to in the shortest possible time.

Our goal is to develop a deep learning model suitable for the classification of conflict information. This model should be able to classify conflict categories and discover category-related topics. Most importantly, the model must have high reliability as the consequences of conflicting information are often very serious. We therefore want to combine text classification, topic modelling and interpretable analysis to solve the problem.

Text classification assigns category labels to different texts for the purpose of distinguishing textual information. Recurrent neural networks (RNNs), convolutional neural networks (CNNs) and graph neural networks (GNNs) have all been applied to text classification tasks (Tai et al., 2015; Zhu et al., 2015; Cheng et al., 2016; Kalchbrenner et al., 2014; Kim, 2014; Johnson and Zhang, 2015; Peng et al., 2018). More recently, Sun et al. (2019) provides a fine-tuned BERT-based pre-training model (Devlin et al., 2019) for text classification tasks generic solution with new state-of-the-art results on eight extensively studied text classification datasets.

The topic model is designed to automatically find a range of topics and topic words from a collection of documents. One of the most classic topic models is latent dirichlet allocation (LDA) (Blei et al., 2003), which is an unsupervised, non-hierarchical model. Many subsequent research has been based on LDA, such as the Hierarchical Latent Dirichlet Allocation (HLDA) proposed by Griffiths et al. (2003). In 2016, Miao et al. (2016) proposed a generative neural variational document model (NVDM), which models the likelihood of documents using a Variational Auto-Encoder (VAE) (Kingma and Welling, 2013). In order to purposefully uncover topic words related to the target (e.g. sentiment), many researchers have also proposed alternative approaches. For example, Ding et al. (2018) added topic consistency to the training as part of the loss as well, thus making the latent variables dependent on the topic target as well.

Neural network-based deep learning can be described as a black box, and humans are not yet able to fully explain or peer into the entire deep learning process. So the question arises whether humans can be trusted with the decision-making mechanisms of such data-driven AI systems. The lack of interpretability leads to a reduction in the reliability of deep learning, hence the importance of interpretable analysis. In an earlier study, Koh and Liang (2017) hoped to find parts of the training data/training points that could be used as a basis for interpretation by introducing the influence func-

tion. Some researchers, on the other hand, have tried to find explanations for the prediction results from the test data itself. Such explanations can be found by perturbing the data (Li et al., 2016), extracting attention weights (Wiegreffe and Pinter, 2019) or calculating the saliency scores of the input sequences (Jain et al., 2020), etc. Lei et al. (2016); Jain et al. (2020) used a combination of generators and encoders to extract rationales.

2 Preliminary Works

Text classification and topic modelling have been important areas of research in natural language processing. These two areas are extremely interrelated, but few studies have effectively integrated them into a unified system. One successful example is the CANTM model proposed by Song et al. (2021) on topic modelling of online text messages during the Covid-19 epidemic, which is able to effectively identify disinformation related to Covid-19 and simultaneously classify the information, helping to address issues such as citizens' distrust of government and healthcare.

The architecture of CANTM is shown in Figure 1. The model is divided into three parts, BERT embedding, the classifier-regularised VAE (M1) and the classifier-aware VAE (M2), where the VAE architectures are used as topic models. The model first uses a BERT pre-trained model to extract segment embeddings h from the input text sequence x. In the encoder part of M1, h is transformed into the parameters μ and σ of the Gaussian distribution via the linear layers $linear_\mu$ and $linear_\sigma$ respectively. The aim of the M1 encoder is to generate the latent variable z, which can be considered as hidden topics. The M1 decoder part uses the latent variable z as input to reconstruct the bag of words of the input text. The M1 classifier also uses the latent variable z as input, and generates classification probabilities after passing through a fully connected layer containing a softmax activation function. Note that since the classifier uses hidden topics as the basis for classification, it has not seen real data, which can reduce the overfitting of the model. The architecture of M2 is similar to M1, except that it takes the classification probabilities $\hat{y}$ output from M1 as input as well, in order to generate hidden topics z_s guided by the classification information. Furthermore, the M2 classifier is not used to output the final classification, but only to compute joint loss during training. The joint loss function of CANTM is a combination of the loss functions of its subcomponents and is calculated as

$$\mathcal{L} = \lambda\mathcal{L}_{cls} - ELBO_{x_{bow}} - ELBO_{x_{bow},\hat{y}} \\ -\mathbb{E}_{\hat{y}}[log\ p(x_{bow}|\hat{y})] \quad (1)$$

CANTM has good classification and topic discovery capabilities, but it is not fully suitable for conflict information. Firstly, it does not introduce interpretability analysis to demonstrate the reliability of the model. Secondly, the topics discovered by CANTM are to some extent disturbed by a large number of neutral words present in the input text, thus making the relevance of the discovered topic words to the category information reduced. Moreover, the CANTM architecture has redundant parts, which affects its computational efficiency.

3 Methodology

Our model is based on an improvement of CANTM, which we call Classification-Aware Neural Topic Model Combined With Interpretable Analysis (CANTM-IA). CANTM is used as the base model because it combines text classification and topic modelling, which aligns with our goals. Secondly, the stacked VAE architecture of CANTM effectively allows us to discover the hidden topics of the target categories. In addition, topics can also be seen as an interpretation of the classification model, which facilitates our interpretability analysis and improvement of the model in conjunction with rationale.

We introduced interpretability analysis specifically by calculating the attention weights of the last layer in the BERT pre-trained model corresponding to the CLS labels and averaging them into the saliency score of the corresponding word piece. The magnitude of the saliency score is used as a visual representation of the importance of different parts of the original sample, and the parts with high saliency scores are used as the rationales of the sample. BERT parameters are frozen during training and only the last transformer encoding layer weights are unlocked for fine-tuning.

Afterwards, we use the saliency scores of the rationales instead of the bag of words of the entire input sequence as the reconstruction target in the VAE architecture. This has several advantages. First, using rationales (the part of the input sequence with high contribution) as the reconstruction target allows the topic model to focus more on the important information of the input
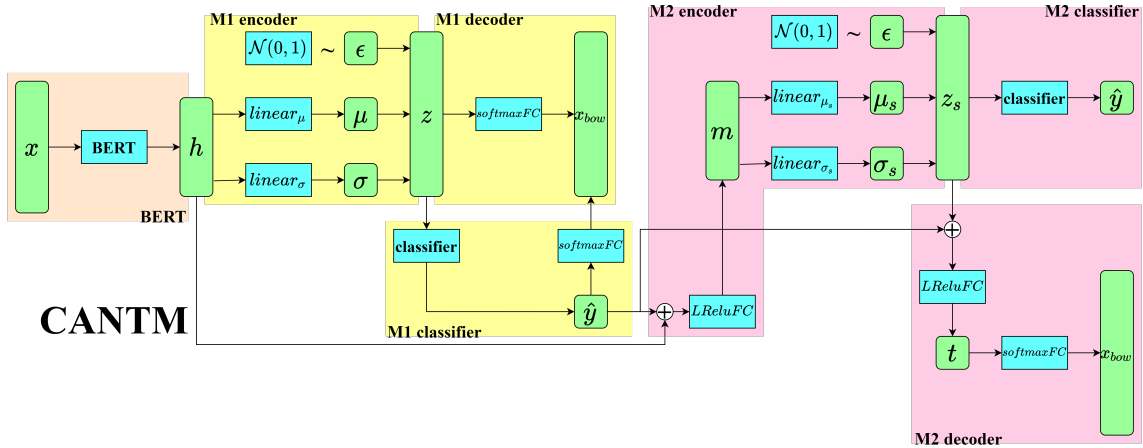
Figure 1: The architecture of the CANTM.

sequence, which can reduce the interference of category-irrelevant words by the topic words and indirectly improve the classification performance of the model. Second, since the decoder uses rationales to guide the discovery of hidden topics and the classifier uses hidden topics for classification, it can be argued that these rationales explain both the hidden topics and the classification results.

In addition, there is a redundant structure in the M2 decoder part of the CANTM model. As shown in Figure 1, m, as the variable that combines the input h with the classification result $\hat{y}$, already introduces classification information for the rest of M2. That is, the process of generating the variable z_s has been guided by the classification information, which generates the class-aware topics. Therefore, there is no need to reintroduce the classification result $\hat{y}$ in the decoder part of M2, and the purpose can be achieved by directly reconstructing the target using z_s as the hidden topic variable.

Combining the above optimisation methods, the modified CANTM-IA model is shown in Figure 2.

4 Experiments

4.1 Dataset

We use The Armed Conflict Location & Event Data Project (ACLED), a disaggregated data collection, analysis, and crisis mapping project, as our source dataset (Raleigh et al., 2010). The ACLED dataset collects six types of events. We use data spanning a full 3 years between 25 June 2019 and 24 June 2022 as experimental data. Of these, the volume of data for the conflict category Protests is 415,588, which far exceeds the volume of data for the other categories. In order to ensure a balanced dataset,

a quarter of the data, i.e. 103,897 items, are randomly selected as the data of category Protests for the experiment. In addition, 50,000 texts from WMT News Crawl Dataset [1] are used as the out-of-domain data. The details of the experimental dataset are shown in table 1, with 90.43% of the ACLED data and 9.57% of the regular news data. The training set, validation set and test set are sampled from the original dataset in a 7:1:2 ratio

4.2 Experimental Setup

We compare our CANTM-IA model with two strong baseline models: **BERT** and **CANTM**. For BERT model, a linear layer of dimension 300 is connected to BERT [CLS] Token output and uses a fully-connected layer with a softmax activation function as a classifier to output the classification results. For CANTM model, we using a bag of words of size 500 and a hidden topic variable of dimension 100.

Three sets of experiments are conducted to compare the choice of parameters and the impact on **CANTM-IA**. The first set uses rationales with a ratio of 10% of the number of tokens in the input text as the reconstruction target, denoted as CANTM-IA (ratio 0.1). In the second set, this proportion is 50% and is denoted as CANTM-IA (ratio 0.5). In addition, a fine-tuning experiment is carried out to fine-tune the model parameters using the CANTM-IA architecture on the trained CANTM model for only 1 epoch. The rationales used for the fine-tuning experiment are scaled to 50% and the model is denoted as CANTM-IA (fine-tune). Other model parameters are kept consistent with
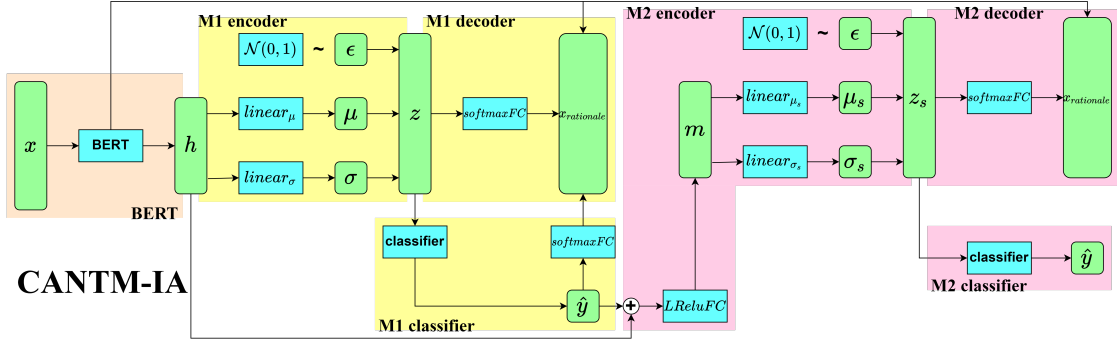
Figure 2: The architecture of the CANTM-IA.

Type of conflict	Battles	Explosions/ Remote violence	Protests	Riots	Strategic developments	Violence agai- nst civilians	Out of domain	Total
Train	76202	61222	72727	34097	30294	56329	35000	365871
Valid	10887	8747	10390	4872	4328	8047	5000	52271
Test	21772	17493	20780	9744	8656	16094	10000	104539
Total	108861	87462	103897	48713	43278	80470	50000	522681

Table 1: Information of the experimental data set.

Model	Accuracy	F-1
BERT (baseline)	0.9738	0.9749
CANTM (baseline)	0.9751	0.9760
CANTM-IA (fine-tune)	0.9766	0.9775
CANTM-IA (ratio 0.1)	0.9774	0.9787
CANTM-IA (ratio 0.5)	**0.9780**	**0.9791**

Table 2: Comparison of the classification performance.

the CANTM baseline system.

We use BERT-base-uncased in experiments, only the last transformer encoding layer is unlocked for fine-tuning, and remaining BERT parameters are frozen during training.

4.3 Results

The overall classification results are shown in Table 2. BERT is a strong baseline with a solid classification accuracy (0.9738). On this basis, CANTM and CANTM-IA still obtained better classification performance by using hidden topics as the basis for classification. The best performing CANTM-IA (ratio 0.5) model achieved an accuracy of 0.9780 and an F1 score of 0.9791, which demonstrates the effectiveness of using hidden topics as a basis for classification. Furthermore, the classification performance of the CANTM-IA (fine-tune) is improved over the CANTM model, even after only 1 fine-tuning. This suggests a positive contribution of the topic model guided by rationale to the effectiveness of text classification. The F1 scores for each sub-category in the dataset are given in Table 3.

It should be noted that since the ACLED data is cleaned by a professional data agency, the content of the data is to a large extent highly normative and accurate. As a result, classification performance can be extremely good even for the baseline model. This makes it appear that the improved model cannot outperform the baseline model by much in terms of experimental results. However, in this case, due to the large amount of data in the dataset, even a subtle advantage is evident in the face of the number of accurate predictions.

We show the top 10 topic words for each conflict category for the CANTM and CANTM-IA (ratio 0.5) models in table 4. As can be seen, the category-related topic words extracted by CANTM already provide a good overview for each conflict category. However, there are still many neutral words such as 'report', 'city' and 'unknown' in the CANTM results. This is due to the fact that CANTM uses a bag of words from the complete input sequence for topic reconstruction, which makes a large number of neutral words that appear in the conflict text influential in the reconstruction process. The weights of the reconstruction matrix with respect to this token is strengthened during training, and the relevance of this word to the relevant category is increased. In contrast, CANTM-IA cleverly reduces the influence of such neutral words. Because CANTM-IA uses rationales as the reconstruction target, this allows the model to focus more on the conflict-related information itself and thus ignore irrelevant neutral words. This results in a greater concentration of topic words that are relevant to the classification results and more representative of the categories. It also demonstrates the effectiveness of

Model	Battles	Explosions/Remote violence	Protests	Riots	Strategic developments	Violence against civilians	Out of domain
BERT (baseline)	0.9583	0.9817	0.9904	0.9754	0.9693	0.9501	0.9994
CANTM (baseline)	0.9628	0.9836	0.9879	0.9707	0.9736	0.9540	**0.9998**
CANTM-IA (fine-tune)	0.9646	**0.9854**	0.9885	0.9704	0.9769	0.9575	0.9996
CANTM-IA (ratio 0.1)	0.9633	0.9849	0.9910	0.9769	**0.9777**	0.9570	0.9997
CANTM-IA (ratio 0.5)	**0.9655**	0.9847	**0.9911**	**0.9772**	0.9774	**0.9584**	0.9995

Table 3: F1 scores of the models for different categories of classification results.

Type of conflict	Topic words in CANTM	Topic words in CANTM-IA (ratio 0.5)
Battles	forces military fatalities killed positions clashed militants taliban coded azerbaijan	clashed killed clashes fire attacked clash fired attack small militants
Explosions/ Remote violence	shelled forces fatalities injuries unknown positions artillery smm osce airstrikes	shelled casualties fired targeted fatalities total airstrikes artillery killed carried
Protests	report protest people city government members held workers gathered protested	protest protested demonstrated held protesters gathered staged demonstration workers demanding
Riots	report police rioters demonstrators demonstration clashed group people stones injured	rioters demonstrators clashed demonstration set attacked beaten beat burning fire
Strategic developments	property destruction forces military arrested township district seized movement security	arrested set destroyed looted fire seized military destruction burned forces
Violence against civilians	killed shot man men fatality armed found colonia unidentified body	killed shot man attacked armed people found beat abducted dead

Table 4: Top 10 topic words for each conflict type.

Model	Note of event (with highlighted rationales)
BERT (baseline)	Property destruction: Around 13 May 2022 (as reported), in Ta Tang Ku village of Pinlaung township (coded as Pinlaung) (Pa-O Self-Administered Zone, Shan-South state), the PNO soldiers destroyed the Catholic statue of the Virgin Mary and a Catholic church. The statue is well respected.
CANTM (baseline)	Property destruction: Around 13 May 2022 (as reported), in Ta Tang Ku village of Pinlaung township (coded as Pinlaung) (Pa-O Self-Administered Zone, Shan-South state), the PNO soldiers destroyed the Catholic statue of the Virgin Mary and a Catholic church. The statue is well respected.
CANTM-IA (ratio 0.5)	Property destruction: Around 13 May 2022 (as reported), in Ta Tang Ku village of Pinlaung township (coded as Pinlaung) (Pa-O Self-Administered Zone, Shan-South state), the PNO soldiers destroyed the Catholic statue of the Virgin Mary and a Catholic church. The statue is well respected.

Table 5: Comparison of rationales extracted from the same sample. Words in red are rationales.

category-related topic words extraction and ensures the possibility of subsequent analysis of the model.

We show rationale examples in table 5 that are extracted by the different models from the same sample. It can be observed that while the rationales extracted by CANTM from can focus on conflict-related information ("soldiers", "destroyed", "church"), there is also some irrelevant information that is focused on ("as", "reported", "of", "zone"). CANTM-IA (ratio 0.5), on the other hand, focuses precisely and intently on the conflict information itself ("the", "soldiers", "destroyed", "a", "church"). Note that although the words "the" and "a" appear to be meaningless and category-independent words on their own. However, the model incorporates contextual information. Therefore, it can be argued that the CANTM-IA model also combines and pays some attention to coherent semantics, which makes the rationales extracted by CANTM-IA more coherent than previous rationales, and allows for better interpretation of the model's classification decisions and topic selection. The rationales comparison experiment shows that the rationales extracted by CANTM-IA focuses on the conflict information itself and can reasonably and effectively explain the model's conflict type classification results. This ensures the reliability of the model's classification decisions and allows CANTM-IA to provide humans with reliable results for further analysis of conflict information to a certain extent.

5 Conclusion

We proposed a Classification-Aware Neural Topic Model (CANTM-IA) for Conflict Information Classification and Topic Discovery in this paper. The classification results and topic models of CANTM-IA can be reliably interpreted using rationales. Also, rationales are introduced into the topic model to improve model performance. Finally, the model architecture has been optimised. Compared to the baseline systems, CANTM-IA has improved predictive performance, reliability and efficiency. Our future work will be to adapt the model to other types of data and to refine the way in which interpretable analysis is introduced.

6 Ethics and Broader Impact Statement

6.1 Ethics

Only publicly available dataset[2] is used in this paper Raleigh et al. (2010). No ethical approval is required for this work.

6.2 Implications

Our work has several potential practical implications:

- Our model outperforms two strong baselines, BERT and CANTM, in terms of predictive performance. It can also serve as a competitive baseline for future research.

- Our explainable neural topic model, CANTM-IA, can be utilized for other NLP downstream tasks, such as stance detection (Mu et al., 2023) and rumor verification (Derczynski et al., 2017), providing **interpretable predictions**.

Acknowledgments

References

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. Long short-term memory-networks for machine reading. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 551–561, Austin, Texas. Association for Computational Linguistics.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, Vancouver, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ran Ding, Ramesh Nallapati, and Bing Xiang. 2018. Coherence-aware neural topic modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 830–836, Brussels, Belgium. Association for Computational Linguistics.

Thomas Griffiths, Michael Jordan, Joshua Tenenbaum, and David Blei. 2003. Hierarchical topic models and the nested chinese restaurant process. In *Advances in Neural Information Processing Systems*, volume 16. MIT Press.

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

Rie Johnson and Tong Zhang. 2015. Effective use of word order for text categorization with convolutional neural networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 103–112, Denver, Colorado. Association for Computational Linguistics.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 655–665, Baltimore, Maryland. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Diederik P Kingma and Max Welling. 2013. Auto-Encoding Variational Bayes. *arXiv e-prints*, page arXiv:1312.6114.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1885–1894. PMLR.

---

[2]https://www.prio.org/misc/Download.
aspx?file=%2fcscw%2frd%2fReplication+
Data%2fReplication+data_Raleigh+et+al+
47(5).zip

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding Neural Networks through Representation Erasure. *arXiv e-prints*, page arXiv:1612.08220.

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1727–1736, New York, New York, USA. PMLR.

Yida Mu, Mali Jin, Charlie Grimshaw, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2023. Vaxxhesitancy: A dataset for studying hesitancy towards covid-19 vaccination on twitter. *arXiv preprint arXiv:2301.06660.*

Hao Peng, Jianxin Li, Yu He, Yaopeng Liu, Mengjiao Bao, Lihong Wang, Yangqiu Song, and Qiang Yang. 2018. Large-scale hierarchical text classification with recursively regularized deep graph-cnn. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 1063–1072, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Clionadh Raleigh, Andrew Linke, Håvard Hegre, and Joakim Karlsen. 2010. Introducing acled: An armed conflict location and event dataset: Special data feature. *Journal of Peace Research*, 47(5):651–660.

Xingyi Song, Johann Petrak, Ye Jiang, Iknoor Singh, Diana Maynard, and Kalina Bontcheva. 2021. Classification aware neural topic model for covid-19 disinformation categorisation. *Public Library of Science (PLoS)*, (2).

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *Chinese Computational Linguistics*, pages 194–206, Cham. Springer International Publishing.

Kai Sheng Tai, Richard Socher, and Christopher D Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566.

Sarah Wiegreffe and Yuval Pinter. 2019. 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)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Xiaodan Zhu, Parinaz Sobihani, and Hongyu Guo. 2015. Long short-term memory over recursive structures. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1604–1612, Lille, France. PMLR.