Prompt-Based Learning for Thread Structure Prediction In Cybersecurity Forums

Anonymous ACL submission

Abstract

With recent trends indicating cyber crimes increasing in both frequency and cost, it is imperative to develop new methods that leverage data-rich hacker forums to assist in combating ever evolving cyber threats. Defining interactions within these forums is critical as it facilitates identifying highly skilled users, which can improve prediction of novel threats and future cyber attacks. We propose a method called Next Paragraph Prediction with Instructional Prompting (NPP-IP) to predict thread structures while grounded on the context around posts.The experimental evaluation shows that our proposed method can predict the thread structure significantly better than existing methods allowing for better social network prediction based on forum interactions.

1 Introduction

Cybercrimes cost trillions of dollars in damages worldwide each year, impacting different sectors of society ranging from national defense to private industry (Fox, 2021). Current trends indicate a considerable rise in cybercrimes in the next several years as hacker tools become more sophisticated and ubiquitous (Morgan, 2016). This is, in part, due to the advent of the dark web, which has given hackers the opportunity to interact, profit, and exchange information on dark web forums (Goel, 2011). Identifying key user interactions within these dark forums can assist in identifying prominent hackers with knowledge of novel threats as well as predicting potential cyber attacks. Thus, the thread structure of a forum becomes important in generating social networks based on user interactions as shown in Figure 1 (Fu et al., 2007).

Unfortunately, most of the hacker forums are unstructured making it difficult to identify user interactions through post replies in an automated manner. Moreover, although many of these dark forums have a rich source of text information in different threads that discuss specific topics, such as malware, virus, illegal items, and other illegal activities, the recent reports indicate that 90% of posts on popular dark web hacking forums are made by those looking to solicit hacker services instead of the hackers themselves (Technologies, 2021; Culafi, 2021). Traditional methods used to define social networks on unstructured forums such as Creator-oriented Network and Last Reply-oriented Network (L'Huillier et al., 2010) are based upon temporal interaction assumptions that do not consider the full context of the user interactions based on the content of the posts. Kashihara et al. previously introduced a powerful deep learning method called Next Paragraph prediction (NPP) designed to define social networks using posts from the Reddit forums (Kashihara et al., 2020). The NPP method outperformed traditional methods as well as BERT's Next Sentence Prediction (NSP) (Devlin et al., 2019) when defining social networks from posts.

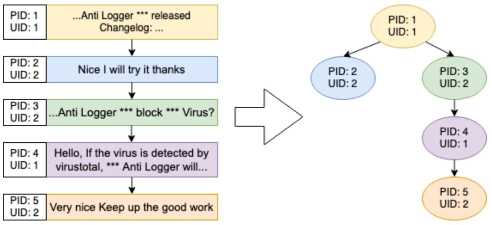
PID: 1
UID: 1
...Anti Logger *** released Changelog: ...
PID: 2
UID: 2
Nice I will try it thanks
PID: 3
UID: 2
...Anti Logger *** block *** Virus?
PID: 4
UID: 1
Hello, If the virus is detected by virustotal, *** Anti Logger will...
PID: 5
UID: 2
Very nice Keep up the good work
PID: 1
UID: 1
PID: 2
UID: 2
PID: 3
UID: 2
PID: 4
UID: 1
PID: 5
UID: 2

Figure 1: Example of a thread from an unstructured Hacker forum in the darkweb (left), and the predicted thread structure (right).

Building upon the NPP method, we propose the Next Paragraph Prediction with Instructional Prompting (NPP-IP) that leverages cutting-edge Prompt-based learning to assist in social network construction from posts. Using a Reddit dataset consisting of over 105 threads and 1,648 posts, we train and evaluate the model against both traditional methods as well as the original NPP method. We also test the model using real unstructured hacker forums data, where 20 threads are manually annotated by human experts to identify interactions based on posts. The results show that NPP-IP performs 2.68 to 4.7 percentage points better than the other existing methods.

Contributions: (i) Prompt-based learning (instructional prompting) is introduced into a deep learning method called Next Paragraph Prediction for social network construction of forum data. (ii) We apply Prompt-based learning to Cybersecurity domain for the first time. (iii) The evaluation results show that our method can predict thread structures better than the existing methods. (iv) The results indicate that the proposed method is robust enough to train using data from one cyber-related forum and apply to another cyber-related forum.

2 Related Work

Thread Structure Prediction: In order to build social networks from forums, member interactions must be correctly identified via posts on threads. There are two network representations introduced (L'Huillier et al., 2010) for building the social network in forums: Creator-oriented Network and Last Reply-oriented Network. The Last Reply-oriented Network is widely used for the social network analysis in the recent works (Phillips et al., 2015; Almukaynizi et al., 2017; Marin et al., 2018; Sarkar et al., 2018; Pete et al., 2020; Johnsen and Franke, 2020). Since these two traditional network conversion approaches are based on limited information and considerable assumptions on interactions between users, the social structures of the networks are likely not accurate representations. Other recent works have predicted helpful posts in the forums (Halder et al., 2019) using a neural network based model that determines whether the post is useful or not. However, the importance of a post has very little utility when predicting interactions and thus social networks.

Instructional Prompting: Building effective discrete prompts for language models (LM) to perform NLP tasks is an active area of research (Schick and Schütze, 2021; Scao and Rush, 2021; Tam et al., 2021; Reynolds and McDonell, 2021). Such prompts are often extremely short and may not include a complete definition of complex tasks. In contrast, the recent works (Mishra et al., 2021a,b) give instructions encode detailed instructions as they were used to collect the dataset. Driven many empirical analysis by (Mishra et al., 2021a), the framing instructional prompting has demonstrated considerable improvements across LMs.

3 Model Description

Our proposed NPP-IP model is based on infusing the original dataset with specific task instructions using an instruction prompting function. Formally, the instruction prompting function $f_{prompt}(\cdot)$ is defined as

$$f_{prompt}(x) = I||x, \tag{1}$$

where || represents concatenation of instruction prompt I with training sample x. Instruction prompt I is formally defined as:

Task Description:
You are given two posts and you need to generate True if they are the direct reply relation, otherwise generate False.
Positive Example:
post1: Windows Defender Gets a New Name: Microsoft Defender
post2: Bring back MSE and its ui even logo looks cool...
output: True
Negative Example:
post1: Windows Defender Gets a New Name: Microsoft Defender
post2: Title says it
output: False"

Training sample x is formally defined as

$$x = \text{Post } k \;||\; \text{[sep]} \;||\; \text{Post } k+i, \tag{2}$$

which represents a pair of concatenated posts at index k and $k+i$ with a separation key $[sep]$, such that $i \neq 0$.

The NPP-IP model leverages five framing techniques defined in (Mishra et al., 2021a) for framing the instruction prompting information I. First, the **Use Low Level Patterns** technique is accomplished by providing a simple task descriptor to correctly output a value of True or False if a reply relationship exists between posts without including any cybersecurity jargon. Second, **Itemized Instructions** are provided via the positive and negative examples with the corresponding output in bulleted list format for thread structure prediction. The positive and negative examples also fulfill the **Break It Down** technique by defining simpler subtasks corresponding to identifying negative and positive examples. This is also where cybersecurity information is introduced into the instructional prompt. Next, **Enforce Constraints** is accomplished by constraining the examples to their

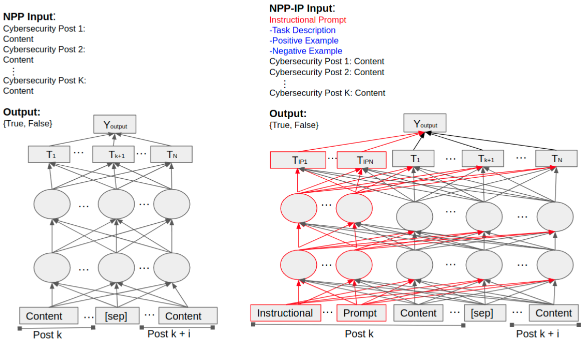

Figure 2: The original NPP model (Left) combines a pair of posts to predict whether one post is a response to the other. Our NPP-IP model (Right) incorporates instruction prompt information into the NPP structure allowing for task information to be leveraged.

respective outputs of True or False. Lastly, the **Specialize Instructions** technique is accomplished by specifically stating the expected output in both task description and examples.

Figure 2 shows the BERT-based neural network structure used by an NPP model as well as the resultant NPP-IP model after introducing instructional prompting information. The original dataset gives two posts as its input, where the label space is defined as {True, False}, defining whether posts share a direct response relation or not. Including instructional prompting provides critical task information for both positive and negative cases, which are then used in the embedding and subsequent prediction task during training.

4 Evaluation and Results

Datasets: A curated Reddit dataset (Kashihara et al., 2020) was used to train and evaluate our proposed model. The Reddit dataset is ideally suited for thread structure analysis given its tree-like structure within different threads. Our proposed model was also evaluated using 20 hacker forum threads from three English hacker forums annotated by human experts, which is referred to as the "Hacker Forums" dataset. The forum thread data is from CYR3CON[1]. The average posts per thread is 15.4. Four cybersecurity experts checked posts in each thread, and annotated a relation of two posts in a thread which the two posts are direct response relations or not. The site names and usernames are anonymized. The topic, thread, and post information from the Reddit and Hacker Forums dataset are provided at A.1.

Method	Model	P	R	F1
CO	-	0.00	1.00	0.01
LR	-	0.72	0.12	0.20
NPP	BE-B	0.42	0.46	0.44
	BE-L	0.36	0.51	0.42
	RB-B	0.59	0.33	0.43
	RB-L	0.41	0.58	**0.48**
NPP-IP	BE-B	0.48	0.46	**0.47**
	BE-L	0.64	0.41	**0.50**
	RB-B	0.62	0.43	**0.51**
	RB-L	0.39	0.56	0.46

Table 1: Results from the Reddit test data show that the NPP-IP method outperformed all other methods for thread structure prediction across all but one of the different BERT language models analyzed.

Metrics and Task: Our proposed NPP-IP method was evaluated against several different methods for thread structure prediction using cybersecurity related posts. Two language models, BERT (BE) and RoBERTa (RB), were explored when training the NPP and proposed NPP-IP models, where -B and -L represent base and large models for each LM respectively. As shown in Figure 3, our NPP-IP method outperformed the original NPP method based on the F1 score using the Reddit data across all but one of the LMs.

We compared performance with well known methods, Creator-Oriented Network (CO) and Last Reply-Oriented Network (LR) using Precision (P), Recall (R), and F1 score (F1) metrics reported in Tables 1 and 2 for Reddit and Hacker Forums datasets, respectively. Both tables show a clear improvement from our NPP-IP method compared to most methods across both datasets. As shown in Table 2, in two of the three hacker forums, our proposed NPP-IP method with BERT-B LM reached the highest F1 score while NPP-IP with BERT-L LM recorded the highest F1 score for the third forum.

Libraries and Hyperparameters: In order to build, train, and evaluate both NPP and NPP-IP methods, we use publicly available libraries and set hyperparameters, and they are described in A.2.

5 Analysis and Discussion

As far as the authors are aware, this is the first time that instructional prompts have been applied to text-based cybersecurity data. As the results of Reddit and Hacker Forums datasets show in Table 1 and Table 2, NPP-IP performs better than NPP in most of the cases. The improvement on the Reddit dataset from NPP to NPP-IP across different LMs, ranged between $3 - 8\%$ across F1 scores. Similar

[1] https://www.cyr3con.ai

		Forum1			Forum2			Forum3		
Method	Model	P	R	F1	P	R	F1	P	R	F1
CO	-	0.31	1.00	0.47	0.27	1.00	0.43	0.12	1.00	0.21
LR	-	0.50	0.00	0.01	0.50	0.09	0.16	0.50	0.13	0.21
NPP	BE-B	0.40	0.37	0.39	0.37	0.61	0.46	0.33	0.65	0.44
	BE-L	0.94	0.27	0.41	0.50	0.34	0.41	0.50	0.33	0.40
	RB-B	0.59	0.38	0.46	0.29	0.50	0.37	0.48	0.43	0.41
	RB-L	0.54	0.55	0.54	0.55	0.58	0.54	0.45	0.40	0.41
NPP-IP	BE-B	0.55	0.39	0.45	0.71	0.63	**0.67**	0.61	0.56	**0.58**
	BE-L	0.70	0.43	**0.53**	0.85	0.31	0.45	0.61	0.60	0.57
	RB-B	0.50	0.37	0.42	0.52	0.58	0.48	0.53	0.84	0.46
	RB-L	0.50	0.87	0.43	0.50	0.34	0.41	0.50	0.33	0.40

Table 2: Results from each of the anonymous hacker forums demonstrated that the NPP-IP model outperformed all other models. The NPP and NPP-IP models were both trained with Reditt data further demonstrating NPP-IPs inference performance robustness on unrelated cyber forums.

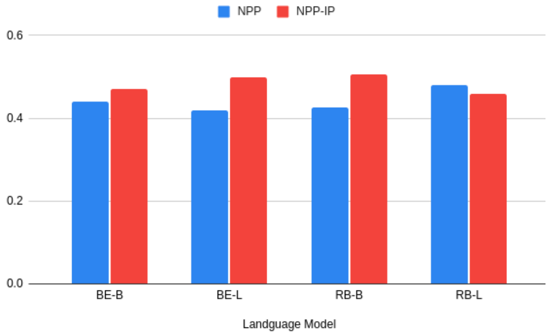

Figure 3: An F1 score comparison of the Reddit Forum data using different BERT-based language models indicates that our proposed NPP-IP model (Red) outperforms NPP (Blue) in all but one language model, achieving the highest score using RoBERTa LM.

improvements were observed using the real world Hackers Forums dataset, ranging between $3 - 11\%$ difference across F1 scores. This is despite the fact that the models were trained on Reddit only data. As shown, in Table 3, the Reddit dataset is comprised of cybersecurity related topics across the different threads. **This evidence is consistent with NPP-IP's ability to better detect and leverage cybersecurity related information compared to other well known methods for social network construction based on thread structure prediction.** Moreover, the framing of the instructional prompt using cyber related information may also have improved its performance across different forums. This is a significant discovery since annotating new datasets, especially in the cyber realm, is costly, requiring considerable human experts' efforts to collect a decent size of data for training and testing. More research needs to be conducted to determine the extent to which framing cyber related instruction prompts can make text-based analysis more robust across different cyber forums and datasets.

As precision and recall scores in Tables 1 and 2 show, both are considerably low in Reddit dataset, with recall scores much lower than precision scores in the Hacker Forums dataset. One possible explanation for this observed behavior is that publicly available pre-trained LMs were used. These LMs are pre-trained by a wide range of topics across a massive size of data. However, the cybersecurity field is in a constant state of flux - changing the meaning of words and adding new words quite frequently. We suspect that LMs could not understand many of the cybersecurity keywords in posts predicting thread structures that were consistent with actual social interaction. Thus, re-training LMs with cybersecurity data should be explored to improve the performance.

6 Conclusion

Predicting thread structures within cybersecurity forums is a crucial component in defining key social networks used to identify prominent users who provide useful information. Identifying these users can facilitate prediction and prevention of future cyber incidents and attacks.

A prompt-based learning model called Next Paragraph Prediction with Instructional Prompting (NPP-IP) for predicting thread structures across different cybersecruity topics was introduced. The method was evaluated using two different datasets and compared against several well known methods. The results show that the NPP-IP method had considerable improvement over existing methods, achieving the highest F1 score across different real world hacker forum datasets.

Ethical Considerations

In this research, we use one dataset from the other work with the agreement of using the dataset for this research only. We created the Hacker Forum dataset where the data from CYR3CON, and they already anonymized the site names and usernames. We have an agreement with CYR3CON to use the data for this research only, and not sharing the data in public. The Hacker Forums dataset has randomly picked 20 threads that have average 15.4 posts per thread from three English hacker forums. The dataset was annotated by four cybersecurity experts (employees of CYR3CON) in a week as a part of their jobs. Our goal is to construct thread structure from unstructured forums. Further precautions taken include not identifying individuals (including not publishing usernames), and presenting results objectively. In addition, we use well-known publicly released language models, BERT and RoBERTa, for our experiments.

References

Mohammed Almukaynizi, Alexander Grimm, Eric Nunes, Jana Shakarian, and Paulo Shakarian. 2017. Predicting cyber threats through hacker social networks in darkweb and deepweb forums. In *Proceedings of the 2017 International Conference of The Computational Social Science Society of the Americas*, page 12. ACM.

Alexander Culafi. 2021. Ninety percent of dark web hacking forum posts come from buyers. `shorturl.at/fEGW4`. Accessed: 2022-04-01.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

William Falcon and Kyunghyun Cho. 2020. A framework for contrastive self-supervised learning and designing a new approach. *arXiv preprint arXiv:2009.00104*.

Jacob Fox. 2021. Cybersecurity statistics 2021. `https://www.cobalt.io/blog/cybersecurity-statistics-2021`. Accessed: 2022-04-01.

Tianjun Fu, Ahmed Abbasi, and Hsinchun Chen. 2007. Interaction coherence analysis for dark web forums. In *IEEE International Conference on Intelligence and Security Informatics, ISI 2007, New Brunswick, New Jersey, USA, May 23-24, 2007, Proceedings*, pages 342–349.

Sanjay Goel. 2011. Cyberwarfare: connecting the dots in cyber intelligence. *Communications of the ACM*, 54(8):132–140.

Kishaloy Halder, Min-Yen Kan, and Kazunari Sugiyama. 2019. Predicting helpful posts in open-ended discussion forums: A neural architecture. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3148–3157.

Jan William Johnsen and Katrin Franke. 2020. Identifying proficient cybercriminals through text and network analysis. In *IEEE International Conference on Intelligence and Security Informatics, ISI 2020, Arlington, VA, USA, November 9-10, 2020*, pages 1–7. IEEE.

Kazuaki Kashihara, Jana Shakarian, and Chitta Baral. 2020. Social structure construction from the forums using interaction coherence. In *Proceedings of the Future Technologies Conference*, pages 830–843. Springer.

Gaston L'Huillier, Héctor Álvarez, Sebastián A. Ríos, and Felipe Aguilera. 2010. Topic-based social network analysis for virtual communities of interests in the dark web. *SIGKDD Explorations*, 12(2):66–73.

Ericsson Marin, Jana Shakarian, and Paulo Shakarian. 2018. Mining key-hackers on darkweb forums. In *1st International Conference on Data Intelligence and Security, ICDIS 2018, South Padre Island, TX, USA, April 8-10, 2018*, pages 73–80.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. 2021a. Reframing instructional prompts to gptk's language. *CoRR*, abs/2109.07830.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2021b. Natural instructions: Benchmarking generalization to new tasks from natural language instructions. *CoRR*, abs/2104.08773.

Steve Morgan. 2016. Hackerpocalypse cybercrime report 2016. `https://cybersecurityventures.com/hackerpocalypse-cybercrime-report-2016/`. Accessed: 2022-04-01.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. In *NIPS-W*.

Ildiko Pete, Jack Hughes, Yi Ting Chua, and Maria Bada. 2020. A social network analysis and comparison of six dark web forums. In *IEEE European Symposium on Security and Privacy Workshops, EuroS&P Workshops 2020, Genoa, Italy, September 7-11, 2020*, pages 484–493. IEEE.

Elizabeth Phillips, Jason RC Nurse, Michael Goldsmith, and Sadie Creese. 2015. Extracting social structure from darkweb forums.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama Japan, May 8-13, 2021, Extended Abstracts*, pages 314:1–314:7. ACM.

Soumajyoti Sarkar, Mohammad Almukaynizi, Jana Shakarian, and Paulo Shakarian. 2018. Predicting enterprise cyber incidents using social network analysis on the darkweb hacker forums. In *2018 International Conference on Cyber Conflict, CyCon U.S. 2018, Washington, DC, USA, November 14-15, 2018*, pages 1–7.

Teven Le Scao and Alexander M. Rush. 2021. How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 2627–2636. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Few-shot text generation with natural language instructions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 390–402. Association for Computational Linguistics.

Derek Tam, Rakesh R. Menon, Mohit Bansal, Shashank Srivastava, and Colin Raffel. 2021. Improving and simplifying pattern exploiting training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 4980–4991. Association for Computational Linguistics.

Positive Technologies. 2021. Custom hacking services. `https://www.ptsecurity.com/ww-en/analytics/custom-hacking-services/`. Accessed: 2022-04-01.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics.

A Appendix

A.1 Dataset Statistics

Table 3 and Table 4 show the basic statistics of Reddit and Hacker Forums dataset respectively.

Topic Name	TH	Posts
cyber_security	8	48
AskNetsec	14	338
ComputerSecurity	12	110
cyberpunk	11	176
cybersecurity	11	158
Hacking	12	370
Hacking_Tutorial	12	110
Malware	9	82
Malwarebytes	8	72
security	8	184

Table 3: The Reddit dataset consisted of ten cybersecurity topics. “TH" is defined as the number of threads in each topic while “Posts" is defined as the number of Posts across the different threads.

Forum #	TH	Posts
Forum1	7	169
Forum2	7	80
Forum3	6	58

Table 4: The Hacker Forums dataset consisted of 20 threads from three hacker forums. “TH" is defined as the number of threads in each topic while “Posts" is defined as the number of Posts across the different threads.

A.2 Libraries and Hyperparameters

In order to build, train, and evaluate both NPP and NPP-IP methods, torchtext 0.8.0 and PyTorch 1.7.1 (Paszke et al., 2017), pytorch-lightning 1.2.2 (Falcon and Cho, 2020), and transformers 3.4 (Wolf et al., 2020) on Google Colab (Nvidia K80 12 GB GPU) were used. We use the hidden dropout probability as 0.15. The batch size was set to 8 and the learning rate was set to 5e-6. The model was trained with $>$ 10 epochs. Convergence was observed around 3 epochs with limited overfitting and the maximum sequence length was set to 250.