A APPENDIX

A.1 Named Entity Recognition

We use named entity recognition (NER) for tweet normalization for tweet classification in Section 2.2.2. NER is a subtask of information extraction that seeks to locate and classify named entity in unstructured text. Examples of named entities in cybersecurity domain are malware, product, company, threat actor, and vulnerability. Since the existing NER tools like NLTK, Stanford CoreNLP, and twitter_nlp are not trained in cybersecurity domain, they cannot be directly applicable to security-related tweets or articles. Thus, we trained our own NER in cybersecurity domain. We fine-tuned the Bert model using 80,000 tweets about cybersecurity events where entities are annotated with company, malware, organization, product, vulnerability, technology, and attack vector. Specifically, a classification layer is added to the pre-trained Bert model, and then all parameters are jointly fine-tuned on NER task using the annotated tweets. We evaluated our Bert-based NER model on 20,000 tweets where entities are also annotated. The precision and recall of malware entity were 0.89 and 0.92. Here malware entities include malware family names, botnet names, and exploit tools. Examples are Mirai, WannaCry, and Trickbot. The evaluation results of other entities tagging by our Bert-based NER model are presented in Table 12. We note that we evaluated a vanilla BiLSTM-CRF sequence labeling model with Flair embedding [34] as well, but its performance was worse than the Bert-based NER model. We also note that we hired 10 part-time workers having knowledge of cybersecurity for manual entity annotation of 10,000 tweets for NER model training and testing.

Table 12: Bert-based NER evaluation results

Entity Type	Precision	Recall	F1-score	Entity Type	Precision	Recall	F1-score
Company	0.88	0.88	0.88	Vulnerability	0.98	1.00	0.99
Malware	0.89	0.92	0.91	Technology	0.84	0.92	0.88
Organization	0.83	0.83	0.83	Attack vector	0.93	0.97	0.95
Product	0.89	0.90	0.89				