Detecting Cybersecurity Events from Noisy Short Text

Semih Yagcioglu, Mehmet Saygin Seyfioglu,*Begum Citamak, Batuhan Bardak
Seren Guldamlasioglu, Azmi Yuksel, Emin Islam Tatli
STM A.Ş., Ankara, Turkey
{syagcioglu, msaygin.seyfioglu, begum.citamak, batuhan.bardak,
sguldamlasioglu, azyuksel, emin.tatli} @stm.com.tr

Abstract

It is very critical to analyze messages shared over social networks for cyber threat intelligence and cyber-crime prevention. In this study, we propose a method that leverages both domain-specific word embeddings and task-specific features to detect cyber security events from tweets. Our model employs a convolutional neural network (CNN) and a long short-term memory (LSTM) recurrent neural network which takes word level meta-embeddings as inputs and incorporates contextual embeddings to classify noisy short text. We collected a new dataset of cyber security related tweets from Twitter and manually annotated a subset of 2K of them. We experimented with this dataset and concluded that the proposed model outperforms both traditional and neural baselines. The results suggest that our method works well for detecting cyber security events from noisy short text.

1 Introduction

Twitter has become a medium where people can share and receive timely messages on about anything. People share facts, opinions, broadcast news and communicate with each other through these messages. Due to the low barrier to tweeting, and growth in mobile device usage, tweets might provide valuable information as people often share instantaneous updates such as the breaking news before even being broadcasted in the newswire *c.f.* Petrović et al. (2010). People also share cyber security events in their tweets such as zero day exploits, ransomwares, data leaks, security breaches, vulnerabilities *etc*. Automatically detecting such events might have various practical applications such as taking the necessary precautions promptly as well as creating self-awareness as illustrated in Fig. 1. Recently, working with the cyber security

---
*Corresponding author.

Dear @AppleSupport, we noticed a *HUGE* security issue at MacOS High Sierra. Anyone can login as "root" with empty password after clicking on login button several times. Are you aware of it @Apple?

Figure 1: A cyber security event. A recently discovered security issue has been reported on Twitter which caught public attention. A security fix has been published right afterward.

related text has garnered a lot of interest in both computer security and natural language processing (NLP) communities (*c.f*. Joshi et al. (2013); Ritter et al. (2015); Roy et al. (2017)). Nevertheless, detecting cyber security events from tweets pose a great challenge, as tweets are noisy and often lack sufficient context to discriminate cyber security events due to length limits. Recently, deep learning methods have shown to be outperforming traditional approaches in several NLP tasks (Chen and Manning, 2014; Bahdanau et al., 2014; Kim, 2014; Hermann et al., 2015). Inspired by this progress, our goal is to detect cyber security events in tweets by learning domain-specific word embeddings and task-specific features using neural architectures. The key contribution of this work is two folds. First, we propose an end-to-end learning system to effectively detect cyber security events from tweets. Second, we propose a noisy short text dataset with annotated cyber security events for unsupervised and supervised learning tasks. To our best knowledge, this will be the first study that incorporates domain-specific meta-embeddings and contextual embeddings for detecting cyber security events.

2 Method

In the subsequent sections, we address the challenges to solve our task. The proposed system overview is illustrated in Fig. 2.
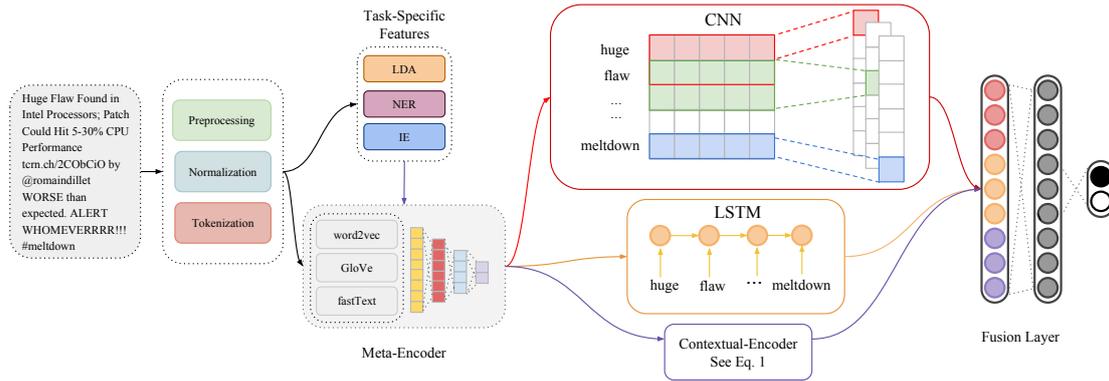
1366

Figure 2: System Overview. A tweet is first pre-processed, then task-specific features and word level meta-embeddings are extracted to represent tokens. Finally, a Bi-LSTM, CNN, and Contextual Encoder are fused to classify the encoded tweet.

2.1 Meta-Embeddings

Word embedding methods might capture different semantic and syntactic features about the same word. To exploit this variety without losing the semantics, we learn meta-embeddings for words.

Word Embeddings. Word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and fast-Text (Joulin et al., 2016; Bojanowski et al., 2016) are trained for learning domain specific word embeddings on the unlabeled tweet corpus.

Meta-Encoder. Inspired by Yin and Schütze (2015) we learn meta-embeddings for words with the aforementioned word embeddings. We use a Convolutional Autoencoder (Masci et al., 2011) for encoding $3xD$ size embeddings to a $1xD$ dimensional latent variable and to reconstruct the original embeddings from this latent variable. Both encoder and decoder are comprised of 2 convolutional layers where 32 neurons are used on each. The encoder part is shown in Fig. 3. We argue that this network learns a much simpler mapping while capturing the semantic and syntactic relations from each of these embeddings, thus leading to a richer word-level representation. Another advantage of learning meta-embeddings for words

is that the proposed architecture alleviates the Out-of-Vocabulary (OOV) embeddings problem, as we still get embeddings from the fastText channel, in contrast to GloVe and word2vec, where no embeddings are available for OOV words.

2.2 Contextual Embeddings

To capture the contextual information, we learn task-specific features from tweets.

LDA. Latent Dirichlet Allocation (LDA) is a generative probabilistic model to discover topics from a collection of documents (Blei et al., 2003). LDA works in an unsupervised manner and learns a finite set of categories from a collection, thus represents documents as mixtures of topics. We train an LDA model to summarize each tweet by using the topic with the maximum likelihood *e.g.* with the topic "vulnerability" for the tweet in Fig 1.

NER. Named Entity Recognition (NER) tags the specified named entities from raw text into predefined categories. Named entities could be more general categories such as people, organizations, or specific entities can be learned by creating a dataset containing specific entity tags. We employ an automatically annotated dataset that contains entities from cyber security domain (Bridges et al., 2013) to train our Conditional Random Field model using handcrafted features, *i.e.*, uni-gram, bi-gram, and gazetteers. The dataset comprises of 850K tokens that contain named entities such as 'Relevant Term', 'Operating System','Hardware', 'Software', 'Vendor', in the standard IOB-tagging format. Our NER model tags "password" as 'Relevant Term' and "Apple" as 'Vendor' for the tweet in Fig 1.

IE. Uncovering entities and the relations between those entities is an important task for detecting cyber security events. In order to address this
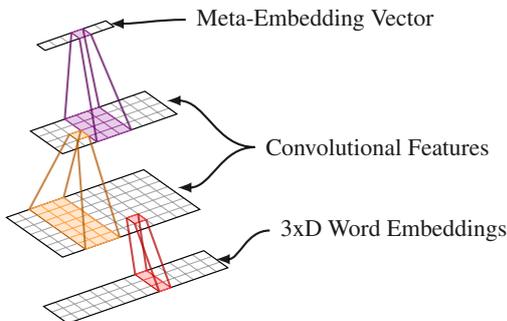

Figure 3: Convolutional encoder as a feature extractor. The decoder is symmetric to the encoder, and in inference time we drop the decoder and use only the encoder network.

we use Information Extraction (IE), in particular OpenIE annotator(Angeli et al., 2015) from the Stanford CoreNLP (Manning et al., 2014). Subsequently, we extract relations between noun phrases with the following dependency triplet $\langle arg_1, rel, arg_2 \rangle$, where arg_1, arg_2 denote the arguments and rel represents an implicit semantic relation between those arguments. Hence, the following triplet is extracted from the tweet in Fig. 1, $\langle we, noticed, huge\ security\ issue \rangle$.

Contextual-Encoder. We use the outputs of LDA, NER and IE algorithms to obtain a combined vector representation using meta-embeddings described in Sec. 2.1. Thus, contextual embeddings are calculated as follows[1].

$$\gamma(\tau) = \frac{f(\varphi(\tau)) + \sum_{i=1}^{N} f(\varsigma(\tau)^i) + \sum_{j=1}^{M} f(\delta(\tau)^j)}{N + M + 1} \quad (1)$$

where γ function extracts contextual embeddings and τ denotes a tweet, f, φ, ς and δ represent meta-embedding, LDA, NER, and IE functions, respectively. Lastly, N and M denote the output tokens.

2.3 Event Detection

Inspired by the visual question answering task (Antol et al., 2015), where different modalities are combined by CNNs and RNNs, we adopt a similar network architecture for our task. Prior to training, and inference we preprocess, normalize and tokenize each tweet as described in Sec. 3.

CNN. We employ a CNN model similar to that of (Kim, 2014) where we feed the network with static meta-embeddings. Our network is comprised of one convolutional layer with varying filter sizes, that is $2, 3, 5$. All tweets are zero padded to the maximum tweet length. We use $ReLU$ as activation and global max pooling at the end of CNN.

RNN. We use a bi-directional LSTM (Hochreiter and Schmidhuber, 1997) and read the input in both directions and concatenate forward and backward hidden states to encode the input as a sequence. Our LSTM model is comprised of a single layer and employs 100 neurons.

3 Experiments

Data Collection. We collected $2.5M$ tweets using the Twitter's streaming API over a period from `2015-01-01` to `2017-12-31` using an initial

set of keywords, henceforth referred as *seed keywords* to retrieve cyber security related tweets. In particular, we use the main group names of cyber security taxonomy described in Le Sceller et al. (2017) as seed keywords e.g. 'denial of service', 'botnet', 'malware', 'vulnerability', 'phishing', 'data breach' to retrieve relevant tweets. Using seed keywords is a practical way to filter out noise considering sparsity of cyber security related tweets in the whole tweet stream. After the initial retrieval, we use `langid.py` (Lui and Baldwin, 2012) to filter out non-English tweets.

Data Preprocessing. We substitute user handles with $mention\$$, and hyperlinks with $url\$$. We remove *emoticons* and reserved keyword RT which denotes retweets. We substitute hashtags by removing the prefix $\#$ character. We limit characters that repeat more than two times, remove capitalization and tokenize tweets using the Twitter tokenizer in `nltk` library. We normalize non-standard forms, *i.e.* writing *cu tmrrw* instead of *see you tomorrow*. Although there are several reasons for that, the most prominent one is that people tend to mimic prosodic effects in speech (Eisenstein, 2013). To overcome this, we use lexical normalization, where we substitute OOV tokens with in-Vocabulary (IV) standard forms, *i.e.* a standard form available in a dictionary. In particular we use UniMelb (Han et al., 2012), UTDallas (Liu et al., 2011) datasets. Lastly, we remove identical tweets and check the validity by removing tweets with less than 3 non-special tokens.

Data Annotation. We instructed cyber security domain experts for manual labelling of the dataset. Annotators are asked to provide a binary label for whether there is a cyber security event in the given tweet or not. Annotators are told to skip tweets if they are unsure about their decisions. Finally, we validated annotations by only accepting annotations if at least 3 among 4 annotators agreed on. Therefore, we presume the quality of attained ground truth labels is dependable. Overall, $2K$ tweets are annotated.

Dataset Statistics. After preprocessing, our initial $2.5M$ tweet dataset is reduced to $1.7M$ tweets where $2K$ of them are labeled[2]. The labeled dataset is somewhat balanced as there are 843 event-related tweets and 1157 non-event tweets. The training and testing sets have 1600 and 400 samples, respectively.

---

[1]We used zero vectors for the non-existent relations.

[2]Available at `http://stmai.github.io/cydec`

Training. We used Keras with Tensorflow back-end in our neural models. For fastText and word2vec embeddings we used Gensim, and for GloVe we used `glove-python` library. For training the word embeddings, we use the entire tweet text corpus and obtain 100 dimensional word embeddings. We set word2vec and fastText model's alpha parameter to 0.025 and window size to 5. For GloVe embedding model, we set the learning rate to 0.01, alpha to 0.75 and maximum count parameter to 100. For embedding models, we determined the minimum count parameter to 5, culminating in the elimination of infrequent words. Consequently, we have 3, 100-dimensional word embedding tensor in which first, second and third channels consist of word2vec, fastText and GloVe embeddings respectively. We then, encode these $3x100$ dimensional embeddings into $1x128$ dimensional representations by using our Meta-Encoder. We train our two channel architecture that combines both LSTM and CNN with 2 inputs: meta-embeddings and contextual embeddings. We use meta-embeddings for feature learning via LSTM and CNN, and their feature maps are concatenated with contextual embeddings in the Fusion Layer. In the end, fully connected layers and a softmax classifier are added, and the whole network is trained to minimize binary cross entropy loss with a learning rate of 0.01 by using the Adam optimizer (Kingma and Ba, 2014).[3]

Baselines. To compare with our results, we implemented the following baselines: `SVM with BoW`: We trained an SVM classifier using Bag-of-words (BoW) which provides a simplified representation of textual data by calculating the occurrence of words in a document. `SVM with meta-embeddings`: We trained an SVM classifier with the aforementioned meta-embeddings. `CNN-Static`: We used Kim (2014)'s approach using word2vec embeddings.

Results. Table 1 summarizes the overall performance of each method. To compare the models, we used four different metrics: accuracy, recall, precision and F1-score. Each reported result is the mean of a 5-fold cross validation experiment. It is clear that our method outperforms various simple and neural baselines. Also, in Table 2, we provide results of our proposed model along with the ground-truth annotations. We also provide results with the different combinations of contextual fea-

Models	Accuracy	Precision	Recall	F1
SVM+BoW	0.75	0.71	0.70	0.70
SVM+Meta-Emcoder	0.71	0.64	0.61	0.63
CNN-static (Yoon Kim, 2014)	0.76	0.72	0.69	0.70
Human	0.65	0.70	0.87	0.59
CNN+Meta-Encoder	0.78	0.78	0.63	0.70
LSTM+Meta-Encoder	0.78	0.74	0.70	0.72
Ours (see Fig. 2)	**0.82**	**0.79**	**0.72**	**0.76**

Table 1: Results

tures, i.e., LDA, NER, IE[4].

Human Study. 8 different subjects are thoroughly instructed about what is considered as a cyber security event and individually asked to label 50 randomly selected tweets from the test set. The results are provided in Table 3.

Error Analysis. In order to understand how our system performs, we randomly select a set of erroneously classified instances from the test dataset.

Type I Errors. Our model identifies this tweet as an event "uk warned following breach in air pollution regulation url" whereas it is clearly about the a breach of a regulation. We hypothesize that this is due to the lack of sufficient training data. Following tweet is also identified as an event "wannacry ransomware ransomwareattack ransomwarewannacry malware url". We suspect that the weights of multiple relevant terms deceive the model.

Type II Errors. Our model fails to identify the following positive sample as an event. For "playstation network was the target of miraibotnet ddos attack guiding tech rss news feed search" our model fails to recognize the 'miraibotnet' from the tweet. We suspect this is due to the lack of hashtag decomposition; otherwise, the model could recognize 'mirai' and 'botnet' as separate words.

Discussions. Cyber security related tweets are complicated and analysing them requires in-depth domain knowledge. Although human subjects are properly instructed, the results of the human study indicate that our task is challenging and humans can hardly discriminate cyber security events amongst cyber security related tweets. To further investigate this, we plan to increase the number of human subjects. One limitation of this study is that we do not consider hyperlinks and user handles which might provide additional information. One particular problem we have not addressed in this work is hashtag decomposition. Error analysis indicates that our model might get confused by challenging examples due to ambiguities and lack of context.

---

[3]See supplementary for hyperparameter choices.

[4]See supplementary for feature combination details.

Tweet	Our Model	GT
that thing where you run corporation phishing test and user does'nt click it but clicks the next message which is real phishing email sigh	0	0
march 03 the fbi investigating alleged data breach at the center for election systems at kennesaw state university	1	1
cia malware codenames are freaking amazing	1	0
proprietary software on malware vista10 is more malicious	0	1
in huge breach of trust deidentified medical history data from millions of australians can be matched to individuals url2	0	1
hackers take aim at your $mention$ account with this new phishing attack cdwsocia	0	1
wannacry ransomware ransomwareattack ransomwarewannacry malware url	1	0

Table 2: Some Example Results

Subjects	Accuracy	Precision	Recall	F1	Cohen's κ
#1	0.62	0.54	1	0.7	0.43
#2	0.54	0.5	0.95	0.65	0.33
#3	0.66	0.58	0.91	0.71	0.42
#4	0.66	0.57	1	0.73	0.46
#5	0.8	0.8	0.73	0.77	0.28
#6	0.66	0.57	0.95	0.72	0.41
#7	0.7	0.63	0.82	0.71	0.31
#8	0.6	0.56	0.60	0.58	0.28
Average	0.65	0.70	0.87	0.59	0.36

Table 3: Human Study Results

4 Related Work

Event detection on Twitter is studied extensively in the literature (Petrović et al., 2010; Sakaki et al., 2010; Weng and Lee, 2011; Ritter et al., 2012; Yuan et al., 2013; Atefeh and Khreich, 2015). Banko et al. (2007) proposed a method to extract relational tuples from web corpus without requiring hand labeled data. Ritter et al. (2012) proposed a method for categorizing events in Twitter. Luo et al. (2015) suggested an approach to infer binary relations produced by open IE systems. Recently, Ritter et al. (2015) introduced the first study to extract event mentions from a raw Twitter stream for event categories DDoS attacks, data breaches, and account hijacking. Chang et al. (2016) proposed an LSTM based approach which learns tweet level features automatically to extract events from tweet mentions. Lately, Le Sceller et al. (2017) proposed a model to detect cyber security events in Twitter which uses a taxonomy and a set of seed keywords to retrieve relevant tweets. Tonon et al. (2017) proposed a method to detect events from Twitter by using semantic analysis. Roy et al. (2017) proposed a method to learn domain-specific word embeddings for sparse cyber security text. Prior art in this direction (Ritter et al., 2015; Chang et al., 2016) focuses on extracting events and in particular predicting the events' posterior given the presence of particular words. Le Sceller et al. (2017); Tonon et al. (2017) focus on detecting cyber security events from Twitter. Our work distinguishes from prior studies as we formulate cyber security event detection problem as a classification task and learn meta-embeddings from domain-specific word embeddings while incorporating task-specific features and employing neural architectures.

5 Conclusion

We introduced a novel neural model that utilizes meta-embeddings learned from domain-specific word embeddings and task-specific features to capture contextual information. We present a unique dataset of cyber security related noisy short text collected from Twitter. The experimental results indicate that the proposed model outperforms the traditional and neural baselines. Possible future research direction might be detecting cyber security related events in different languages.

Acknowledgments

References

Gabor Angeli, Melvin Johnson Premkumar, and Christopher D Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the ACL 2015*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE ICCV*, pages 2425–2433.

Farzindar Atefeh and Wael Khreich. 2015. A survey of techniques for event detection in twitter. *Computational Intelligence*, 31(1):132–164.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. arXiv:1409.0473. Version 7.

Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *IJCAI*, volume 7, pages 2670–2676.

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

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. arXiv:1607.04606. Version 2.

Robert A Bridges, Corinne L Jones, Michael D Iannacone, Kelly M Testa, and John R Goodall. 2013. Automatic labeling for entity extraction in cyber security. arXiv:1308.4941. Version 3.

Ching-Yun Chang, Zhiyang Teng, and Yue Zhang. 2016. Expectation-regulated neural model for event mention extraction. In *HLT-NAACL*, pages 400–410.

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on EMNLP*, pages 740–750.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *HLT-NAACL*, pages 359–369.

Bo Han, Paul Cook, and Timothy Baldwin. 2012. Automatically constructing a normalisation dictionary for microblogs. In *Proceedings of the 2012 joint conference on EMNLP and CoNLL*, pages 421–432. ACL.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in NIPS*, pages 1693–1701.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Arnav Joshi, Ravendar Lal, Tim Finin, and Anupam Joshi. 2013. Extracting cybersecurity related linked data from text. In *Semantic Computing (ICSC), 2013 IEEE Seventh International Conference on*, pages 252–259. IEEE.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. arXiv:1607.01759. Version 3.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. arXiv:1408.5882. Version 2.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv:1412.6980. Version 9.

Quentin Le Sceller, ElMouatez Billah Karbab, Mourad Debbabi, and Farkhund Iqbal. 2017. Sonar: Automatic detection of cyber security events over the twitter stream. In *Proceedings of the 12th International Conference on Availability, Reliability and Security*, page 23. ACM.

Fei Liu, Fuliang Weng, Bingqing Wang, and Yang Liu. 2011. Insertion, deletion, or substitution?: normalizing text messages without pre-categorization nor supervision. In *Proceedings of the 49th Annual Meeting of the ACL: Human Language Technologies: short papers-Volume 2*, pages 71–76. ACL.

Marco Lui and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30. ACL.

Kangqi Luo, Xusheng Luo, and Kenny Qili Zhu. 2015. Inferring binary relation schemas for open information extraction. In *EMNLP*, pages 555–560.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, pages 55–60.

Jonathan Masci, Ueli Meier, Dan Cireşan, and Jürgen Schmidhuber. 2011. Stacked convolutional autoencoders for hierarchical feature extraction. *Artificial Neural Networks and Machine Learning–ICANN 2011*, pages 52–59.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. arXiv:1301.3781. Version 3.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. *EMNLP*.

Saša Petrović, Miles Osborne, and Victor Lavrenko. 2010. Streaming first story detection with application to twitter. In *Human Language Technologies: The 2010 Annual Conference of the NAACL*, pages 181–189. ACL.

Alan Ritter, Oren Etzioni, Sam Clark, et al. 2012. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on KDD*, pages 1104–1112. ACM.

Alan Ritter, Evan Wright, William Casey, and Tom Mitchell. 2015. Weakly supervised extraction of computer security events from twitter. In *Proceedings of the 24th International Conference on World Wide Web*, pages 896–905. International World Wide Web Conferences Steering Committee.

Arpita Roy, Youngja Park, and SHimei Pan. 2017. Learning domain-specific word embeddings from sparse cybersecurity texts. arXiv:1709.07470. Version 1.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM.

Alberto Tonon, Philippe Cudré-Mauroux, Albert Blarer, Vincent Lenders, and Boris Motik. 2017. Armatweet: Detecting events by semantic tweet analysis. In *European Semantic Web Conference*, pages 138–153. Springer.

Jianshu Weng and Bu-Sung Lee. 2011. Event detection in twitter. *ICWSM*, 11:401–408.

Wenpeng Yin and Hinrich Schütze. 2015. Learning meta-embeddings by using ensembles of embedding sets. arXiv:1508.04257. Version 2.

Quan Yuan, Gao Cong, Zongyang Ma, Aixin Sun, and Nadia Magnenat Thalmann. 2013. Who, where, when and what: discover spatio-temporal topics for twitter users. In *Proceedings of the 19th ACM SIGKDD international conference on KDD*, pages 605–613. ACM.