Preventing Poisoning Attacks on AI based Threat Intelligence Systems

Nitika Khurana, Sudip Mittal and Anupam Joshi
University of Maryland, Baltimore County, Baltimore, MD 21250, USA
Email: {nkhur1, smittal1, joshi}@umbc.edu

***Abstract*— As AI systems become more ubiquitous, securing them becomes an emerging challenge. Over the years, with the surge in online social media use and the data available for analysis, AI systems have been built to extract, represent and use this information. The credibility of this information extracted from open sources, however, can often be questionable. Malicious or incorrect information can cause a loss of money, reputation, and resources; and in certain situations, pose a threat to human life. In this paper, we use an ensembled semi-supervised approach to determine the credibility of Reddit posts by estimating their reputation score to ensure the validity of information ingested by AI systems. We demonstrate our approach in the cybersecurity domain, where security analysts utilize these systems to determine possible threats by analyzing the data scattered on social media websites, forums, blogs, etc.**

***Index Terms*— Cybersecurity, Artificial Intelligence, Threat Intelligence, Poisoning Attacks, Credibility**

I. INTRODUCTION

Artificial Intelligence (AI) is widely utilized in diverse domains of industries like, finance, cars, cybersecurity, education, etc. AI systems are 'trained' to learn complex problems and automate them for a larger scale. These systems need training data which is generally extracted and represented in a form that best suits the problem. One such source of data is *overt* or in a traditional cybersecurity sense, a part of the 'Open-source Intelligence' (OSINT) [25]. OSINT includes data from sources such as newspapers, blogs, discussion groups, radio, social media websites, press conferences, journals, technical reports, etc. Online Social Media (OSM) is an OSINT source providing data that is ingested by AI tools working in various fields like finance [15] and cybersecurity [21]. Some of the most commonly used OSM are Twitter, Reddit[1], etc.

In cybersecurity, threat intelligence can be mined using *traditional* sources like NIST's National Vulnerability Database (NVD)[2], United States Computer Emergency Readiness Team (US-CERT)[3], etc. Other sources which are more *non-traditional* are, Twitter, Reddit, blogs, and news. Non-traditional sources are faster than the traditional ones. There is a significant gap between initial vulnerability announcement and NVD release [24]. Vulnerability threat intelligence appears first on non-traditional sources [23]. Mining non-traditional sources is becoming really important. In our previous work, we have developed *CyberTwitter* [21] and *Cyber-All-Intel* [22] systems that mines threat intelligence from various OSINT sources. The systems then represent cybersecurity intelligence in knowledge graphs and vector spaces so it can be used by artificial intelligence based cyber-defense systems.

A new class of 'Analyst Augmentation Systems' are being developed. More security analysts use these Artificial Intelligence based organizational cyber-defense systems to listen for threat intelligence mined from traditional and non-traditional sources, identify new vulnerabilities, analyze network and endpoint activity, find evidence of preplanned attacks and hints of data breaches.

The very 'open' nature of these OSINT sources is its boon and its bane. These open channels are susceptible to 'poisoning attacks' by a malicious entity. In a recent poisoning attack on Twitter, billions were wiped of the US Stock market when an Associated Press tweeted that then President, Barak Obama, had been injured in bomb blasts at the White House. This hack into Associated Press's Twitter account sent Dow Jones plunging 145 points in two minutes and S&P 500 by nearly 1% thereby incurring a loss of $136.5 billion [13]. Data from OSM is vulnerable to misinformation in the form of hoaxes, fake images, videos, and rumors. This traditionally constitutes as fake news [19]. Several of these fake news incidents have caused a loss of money, reputation, infrastructure and in certain cases, threat to human lives.

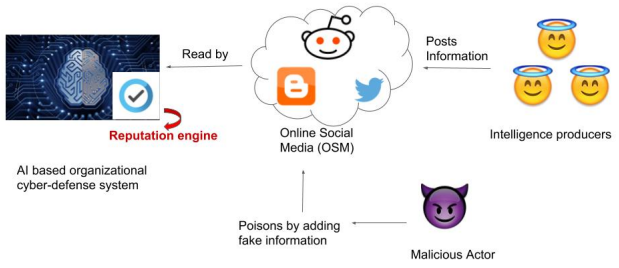

Fig. 1. Attack scenario and proposed defense. Fake or contradictory information added by Attacker is verified using a reputation engine.

Increasing adoption of these non-traditional sources in AI cyber defense systems have created a potential attack surface. In an ideal environment, everything available on non-traditional intelligence sources will be credible and security analysts will mine information from these sources,

[1] https://www.twitter.com, https://www.reddit.com
[2] https://nvd.nist.gov/
[3] https://www.us-cert.gov/

to identify new vulnerabilities and then train their systems [23]. However, in a realistic world, attackers, want to get past these AI cyber defenses by spreading misinformation. They can 'poison' the data by adding incorrect information, for example, an attacker might spread the information that there exists a buffer overflow in Mozilla Firefox, this might trigger a policy change directive by a defensive AI. An attacker might use this as a diversionary tactic against the AI. Figure 1 explains this attack scenario.

They can also put in contradicting information about a valid threat intelligence. for example, an attacker might publish the information that a buffer overflow vulnerability exists in software MySQL, wherein MySQL has a SQL injection vulnerability. In this case the contradicting information will make the AI system more susceptible to an attack. The AI system will devote organizational resources like, analyst time, policy updates, network and endpoint defensive rule updates, etc. to protect against contradictory intelligence. A special case of contradictory information is 'negative' intelligence. For example, in such a scenario the attacker publishes information that there is no SQL injection vulnerability or a valid vulnerability intelligence is false. This will reduce the confidence, that an AI system will place on a valid intelligence.

OSINT as such, if consumed by the AI cyber defense system, can help the attacker evade various security measures thereby putting the organization at risk. Figure 1 explains this attack scenario and proposed defense. In this paper, we ensemble a SVM and an embedding model to build a reputation engine that checks the credibility of gathered intelligence information before it is consumed by the defensive AI. The reputation engine calculates a reputation score for each post and based on the generated score, recommends it for consumption. We use vector embeddings generated using the broad cybersecurity corpus created for the Cyber-All-Intel system [22] (See Section III-C). The reputation score can be used by the AI system and the security analyst to threshold and control the level of trust in the incoming intelligence. The SVM classifier is used to classify posts as 'credible' and 'non-credible' using Reddit posts and 'Redditor' features. More details about our proposed engine are described in Section III.

The remaining paper is organized as follows: Section III describes the background and the related work. Section III discusses our methodology including data collection, vector generation, SVM classification, ensemble generation and reputation score calculation. Section IV summarizes our results. We conclude in Section V.

II. RELATED WORK & BACKGROUND

In this section we discuss the background and the related work in the field of cybersecurity, artificial intelligence, credibility, and provenance.

A. AI for Cybersecurity

Various representation techniques like knowledge graphs and vector space embeddings have been used to provide AI systems with knowledge about cybersecurity.

Knowledge graphs have been used in cybersecurity to combine data and information from multiple sources which then aids a security analyst in her day to day operations. Various ontology based intrusion detection systems [30], [17], [29], [28] have been put forth by researchers. These systems depend on a data repository of system vulnerabilities and threats [16], [21]. These repositories are stored as RDF[4] linked data created from vulnerability descriptions collected from the National Vulnerability Database, Twitter, etc. Joshi et al. [16] extract information on cybersecurity-related entities, concepts and relations which is then represented using custom ontologies for the cybersecurity domain and mapped to objects in the DBpedia knowledge base [10] using DBpedia Spotlight [20]. CyberTwitter [21], a framework to automatically issue cybersecurity vulnerability alerts to users. CyberTwitter converts vulnerability intelligence from tweets to RDF and uses the UCO ontology (Unified Cybersecurity Ontology) [27] to provide their system with cybersecurity domain information. Mittal et al. have also created *CyberAll-Intel* where they have used multiple knowledge representations to store threat intelligence [22].

Systems like the one proposed in [21], [22] that extracts information from OSINT are susceptible to various attacks. For example, a possible attack on our proposed system is that the attacker can 'poison' data sourced through multiple sources like Blogs, Social media, Dark Web, etc. i.e. an attacker can spread the information that there is a vulnerability in Microsoft Windows, even when such a vulnerability does not exist. In such a scenario we need to ensure that the credibility of the information being added to our cybersecurity corpus is checked by a reputation engine as discussed in Section III.

B. Attacks on AI

AI systems are susceptible to threats posed by malicious inputs [23], [24]. Stevens et al. [26] describes how malicious inputs exploiting implementation bugs in ML algorithms pose a threat to organizations. They have defined the term 'poisoning attacks' and 'evasion attacks' as an exploit targeting the training and testing phase respectively. They used a semi-automated technique, called steered fuzzing to explore the attack surface and calculate the magnitude of the threat.

C. Credibility of Intelligence

Several models or tools have been developed over the past to identify 'poisoning' of data in a generic sense. Our work aims at creating a credibility system for Threat Intelligence.

One such system is 'TweetCred' [14], that assigns a 'credibility score' to every tweet to identify fake tweets and thereby providing valuable information during crisis to emergency responders and the public. It was devised to identify the credibility of tweets motivated by false tweets published during 'high impact events'. Rakib et al. used word embeddings on Reddit database based on word2vec skip-gram model to train a random forest classifier to identify

[4] https://www.w3.org/RDF/

cyberbully comments [11]. We build upon these systems to assign a reputation score for threat intelligence mined from Reddit. On Reddit, each account is associated with some meta-data which is the user profile information, the posts written using that account and the network information which comprises of its connections with other user accounts. We use these features and other latent semantic models to compute the reputation score (See Section III).

III. METHODOLOGY

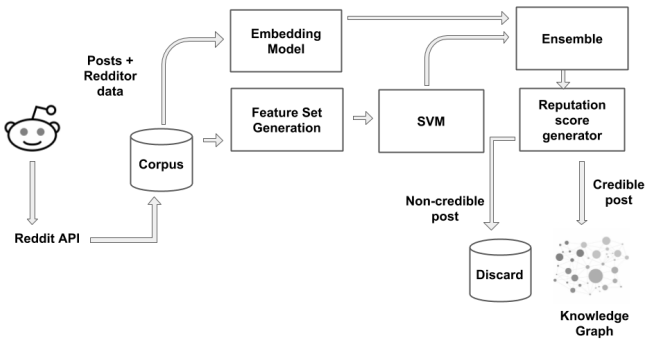

Fig. 2. Architecture of our methodology and analysis.

In this section, we describe the overall architecture (See Figure 2) of our proposed system that includes a reputation engine to calculate the reputation score for each post. The system was created by generating vector embeddings for Reddit posts. We use a semi-supervised learning algorithm, where we have a lot of unlabeled data with a small quantity of labeled data. Our approach leverages the cluster and continuity assumptions which are a universally accepted parts of various semi-supervised learning algorithms [12]. The reputation score is generated using the distance of a post's embeddings from 'credible' and 'non-credible' clusters. The SVM classifier is used to classify posts as 'credible' and 'non-credible' using Reddit posts and 'Redditor' features.

A. Data Collection

Reddit is a social news aggregation, web content rating, and discussion website with over 230 million users [31]. The data is segregated into different tabs within each subreddit. We collected data from Reddit using the Reddit API [5] The API gives an instance of Reddit that can be used to obtain all the 'hot', 'new', 'controversial', 'gilded' or 'top submission' instances. It also provides the data on submitter of the post (also termed as a 'Redditor') and various comments. We collected 14,500 posts corresponding to several cybersecurity subreddits like: cybersecurity, malware, cryptography, cyberlaw and cybersecurityfans, etc.

B. Labeled dataset generation

Human annotators were used to obtain the ground truth for our experiments. From the 14,500 posts, we randomly picked a sample of 2000 posts for annotation. We provided the annotators the definition of credibility and asked them to classify the posts into two classes: 'credible' or 'non-credible'. Annotators were given added information like referred Common Vulnerabilities and Exposures (CVE) database entries and links to verified news websites like, The Washington Post [9], BBC [1], The Guardian [3], CNN [2], Reuters [8], etc. or cybersecurity sources like HackerNews [4], Krebs on Security [5], Microsoft [6], etc. We annotated cyber Reddit posts with the help of 5 graduate students with specialization in cybersecurity to obtain the ground truth regarding the credibility of posts. Each post was annotated by 3 annotators. We calculated the Cohen's Kappa score to check the reliability of the results obtained by annotation. Each post was annotated by at-least 3 annotators to get a good inter annotator agreement. The inter-annotator agreement for all posts was calculated and posts with score > 0.66 were kept. We obtained 1206 posts that served as ground truth with 953 posts entitled as 'credible' and 253 as 'non-credible'.

C. Reddit Post Vectors

In our supervised model, we also incorporated vector projections of the post to help classify them as credible or non-credible. We create embeddings for the posts in which each post is modeled as a 'bag of words' and represented as a sum of it's word embeddings. All the word vectors are summed up to get the total vector value of the post. The word embeddings were taken from the model created by Mittal et al. for their *Cyber-All-Intel* system [22]. To create the Reddit post embeddings we took the following steps:

1) Generate individual cybersecurity word embeddings: We used a cybersecurity corpus collected from multiple OSINT sources like National Vulnerability Datasets, security bulletins, security blogs, Twitter, Reddit, etc. The text corpus and word embeddings were taken from the Cyber-All-Intel system [22]. Taking a corpus collected using different OSINT sources provides the system a more global view of the cybersecurity landscape.
2) Extract cybersecurity concepts and vulnerabilities present in Reddit posts: We use a Security and Vulnerability Concept Extractor (SVCE) to extract terms related to cybersecurity [18], [21]. The SVCE is able to tags every sentence with the following concepts: Means of an attack, Consequence of an attack, affected software, hardware and operating system, version numbers, network related terms, file names and other technical terms.
3) Creating Reddit post vectors: Once we get the output of the SVCE, we fetch the corresponding word embeddings from the word embedding model mentioned in Step 1. Each post is then represented as the sum of cybersecurity term vectors present in that post (This is a slight modification to Doc2Vec[6]). In our imple-

[5] https://www.reddit.com/dev/api/

[6] https://radimrehurek.com/gensim/models/doc2vec.html

mentation we do not include non-cybersecurity terms in the post representation as we empirically found that it adds noise to the system.

Using the ground truth post's vectors we create 2 clusters: 'credible' and 'non-credible'. We use these to compute the reputation score. A visual representation has been shown in Figure 4. We evaluate the quality of vectors generated in Section IV.

D. SVM Classification

We trained a SVM classifier that we ensemble with the embedding model mentioned in Section III-C. We begin by defining the feature set and then train a classifier using Support Vector Machine (SVM). The following are the features that we include in our model, these features have been collected using the Reddit API:

- *Post features*: Length of a post, seconds passed since it was posted, downvotes, upvotes, score, number of comments, number of crossposts and Web of Trust (WOT)[7] values of URLs.
- *Redditor features*: Redditors screen name length, seconds since the user registered, link_karma, comment_karma, verified user email, verified user, user is a moderator or not.

After training Linear SVC on the annotated 1206 posts, we obtained a learned model that classifies posts for credibility. We classified the posts into two classes 'credible' and 'non-credible'. Next we discuss the ensemble model and the reputation score generation. We evaluate all three models in Section IV.

E. Ensemble & Reputation score generation

In our system, we ensemble the two models: SVM classifier and vector embeddings to identify the validity of our posts. We use stacking method to improve the prediction of our system. Stacking models in parallel combines all the classifiers and creates a meta-classifier. In the first step, we utilize the vector embedding model as a base model that is trained on the complete training set of 1206 posts and the post vectors thus obtained as output are collectively used with other identified features for our meta-model, SVM classifier. Figure 3, gives details on how the two classifiers are combined. The embedding model produces a predictive measure Pe and the SVM classifier produces Ps. The weighted sum of these two models gives the final prediction (Pf) for the credibility of a post. The weights We and Ws are learned experimentally. A system analyst can set a threshold for Pf. For our experiments we use $Pf > 0.6$ as credible.

$$Pf = \frac{(We * Pe) + (Ws * Ps)}{(We + Ws)}$$

$$Pf, Ps, Pe \in [0, 1]$$

[7] https://www.mywot.com

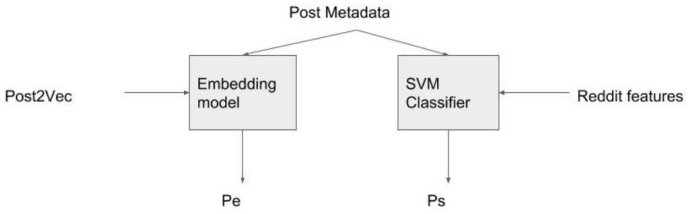

Fig. 3. Ensembling: Stacking Embedding model with SVM Classification.

Next, we wish to create a quantifiable score which can be understood by both the AI system and the security analyst. We calculate the reputation score of a post by determining the distance of the post vector from the cluster centroids created in Section III-C and the output of the Ensemble unit. The score s_c is calculated with respect to the distance from 'credible' cluster (d_c) and the distance from the 'non-credible' cluster (d_i) as:

$$s_c = 1 - \frac{d_c}{d_c + d_i}$$

We use both the ensembled SVM classifier along with the vector embeddings to predict if a post is 'credible' or 'non-credible' and it's reputation score. We also identify the features that serve as strong indicators of credibility for classification by determining the weighted classifier coefficients. We discuss the same in Section IV.

IV. RESULTS & EVALUATION

This section describes the results obtained on classifying posts using the ensembled model. We first evaluate the vector embedding model followed by the SVM model. We also discuss the features that turned out to be strong indicators of credibility.

A. Quality of Vector Embeddings

In Section III-C we discuss our Reddit post vector generation. The post vectors were evaluated manually, by randomly taking 50 posts and then analyzing 5 similar posts retrieved, for each of the 50 using the embedding model. The annotations were done by 2 annotators who evaluated if the retrieved posts were similar to the input post. The mean average precision recorded for the same was 0.59. We would like to point out the fact that this evaluation scheme is really expensive. The results of the credible and non-credible classification using just the embedding model has been shown in Table II.

B. SVM Classification Analysis

We used the Support Vector Machine (SVM) over the selected features described in Section III-D to estimate the credibility of the posts. After training SVM on the annotated 1206 posts, we obtained a learned model that classifies posts for credibility. The results of the credible and non-credible classification using just the SVM model has been shown in Table II.

As a result of our analysis, we identified the following features as strong indicators of credibility: the time at which the post was submitted, the Web Of Trust (WOT) score of the URL in the post, post's length and 'Redditor' features such as link and comment karma. High value of the WOT score of the post URL indicates high credibility of the URL from which the data is extracted. High WOT score websites are observed to be the verified news websites like The Washington Post [9], BBC [1], The Guardian [3], CNN [2], Reuters [8], etc. or cybersecurity sources like HackerNews [4], Krebs on Security [5], Microsoft [6], etc. Thus, presence of a URL in a post showed a strong positive correlation with credibility. The length and submission time of the post and also suggested high credibility of the post; informative and older posts seem to be credible. Some other important indicators were Redditor's link and comment karma. A link karma shows the number of links posted by a 'Redditor' and comment karma exhibits the number of posted comments and upvoted by other 'Redditors'. 'Redditors' who have been active and posted more comments and links are trusted and usually post credible posts. Hence, the post attributes and 'Redditor' features played an important role in determining credibility.

C. Ensembled Model

For the ensembled model, we take the output of the embedding model and the SVM to create a stacked meta-classifier (see Section III-E). On evaluating the ensembled model, we get a ten-fold cross validated accuracy of 71.54%.

	True Positive	**False Positive**
False Negative	188	67
True Negative	77	174

TABLE I

CONFUSION MATRIX FOR BALANCED SET OF 'CREDIBLE' AND 'NON-CREDIBLE' POSTS FOR THE ENSEMBLED MODEL

Metrics	**Ensembled**	**Embedding**	**SVM**
Accuracy	71.541%	66.919%	58.02%
Precision	0.72199	0.66900	0.57
Recall	0.69323	0.62	0.554
True Negative Rate	0.73725	0.6832	0.604
False Positive Rate	0.26274	0.3111	0.395
F1 Score	0.70732	0.6525	0.575

TABLE II

CONFUSION MATRIX AND DERIVED METRICS FOR A BALANCED SET. THE WEIGHTS FOR THE ENSEMBLED MODEL WERE $We = 0.58$ AND $Ws = 0.47$. WE TAKE $Pf > 0.6$ AS CREDIBLE

Table I describes the confusion matrix obtained for the predicted posts. Thus, our analysis for credibility predicted results with an accuracy of 71.541%. We also computed other derived metrics (Table II) for a balanced set of 'credible' and 'non-credible' posts.

We also calculated the reputation score of the posts using their relative distances from the credible and non-credible clusters obtained from ground truth post vectors. Figure 4, shows that posts identified as 'credible' by classification tend to lie in close proximity of credible cluster and 'non-credible' posts lie close to non-credible cluster. The distance from the centroids of the two clusters for two sample posts is listed in Table III. The first post was identified as 'credible' by our analysis and was closer to the credible cluster and the second post was closer to the incredible cluster and identified as 'non-credible'. The minimum of the distances of the post's vector from the centroids of the two clusters gave its reputation score. Hence, post 1 had a reputation score of 0.791 and post 2 received a score of 0.103.

Post	**Distance from Credible cluster**	**Distance from Non-credible cluster**	**Rep. score**
I have just tried it and this exploit just works !!! Joomla powered websites that have "Joomanager 2.0.0"	0.00697	0.02646	0.791
Turns out the Verge fiasco is worse than thought. Devs now having to issue new wallets having accidentally hardforked their own currency trying to fix the attack. Popcorn, salt and GODL overflowing	0.02986	0.00343	0.103

TABLE III

DISTANCE OF POST'S VECTOR FROM CENTROID OF TWO CLUSTERS.

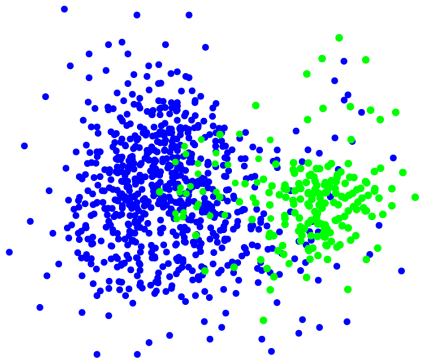

Fig. 4. Visualization of post clusters using t-SNE. Blue cluster represents 'credible' posts and green represents 'non-credible' annotated posts.

V. CONCLUSION AND FUTURE WORK

With the rise in use of online social media (OSM) and data analysis, AI systems have been widely used for predictive analysis. The information extracted from these sources is prone to poisoning.

In the domain of cybersecurity, OSMs have become a source of threat intelligence gathering. This threat intelligence is usually ingested by various cyber-defense systems. The AI systems are exposed to poisoning attacks if we do not perform a credibility check before an intelligence is ingested by a cyber-defense AI. In this paper, we create a reputation

engine to calculate the credibility of the threat intelligence. We have evaluated the credibility of Reddit posts that belong to cybersecurity, cyber, malware, cryptocurrency, cryptomarkets, cyberlaw, etc. subreddits. We created a ensembled semi-supervised model to calculate the reputation of Reddit posts, related to cybersecurity. We ensembled an embedding model and a SVM model. Ground truth was established using manual annotation of posts that were used to train our model and predict the credibility of posts. We classified the posts as 'credible' or 'non-credible' with an accuracy of 71.73%. The reputation score of the posts was evaluated based on the distance of the post vector from the centroids of the clusters plotted for posts in a vector space. We established that both content and 'Redditor' features play a vital role in determining the credibility of a Reddit post.

In the future, we would establish more ground truth data for our analysis to further improve the accuracy of our system. We have used an ensembled semi-supervised approach; such an approach usually yields better results with more annotated data. Getting more annotated data is expensive, access to more ground truth will help in better evaluation and training. Also, we would like to incorporate other online social networks like Quora [7], Twitter, dark web, etc. as they are widely used for discussions about cybersecurity threats and vulnerabilities. We would also like to include a validation scheme where vendors can put their threat intelligence as verified. Vendors can tag their intelligence as verified in the form of a tag or an attribute. We would also like to develop a User Interface or a tag with each post displaying its reputation score or ask for a feedback if the user does not agree with the calculated score.

ACKNOWLEDGEMENT

The work was partially supported by a gift from IBM Research, USA.

REFERENCES

[1] Bbc: Cybersecurity. http://www.bbc.com/news/topics/cz4pr2gd85qt/cyber-security.

[2] Cnn: Cybersecurity. https://www.cnn.com/specials/tech/cybersecurity.

[3] The guardian: Technology. https://www.theguardian.com/technology.

[4] The hacker news. https://thehackernews.com/.

[5] Krebs on security. https://krebsonsecurity.com/.

[6] Microsoft: Cybersecurity. https://www.microsoft.com/en-us/security/default.aspx.

[7] Quora. https://www.quora.com/.

[8] Reuters: Cybersecurity. https://www.reuters.com/subjects/cybersecurity.

[9] The washington post: Cybersecurity. https://www.reuters.com/subjects/cybersecurity.

[10] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. *DBpedia: A Nucleus for a Web of Open Data*. Springer, 2007.

[11] Tazeek Bin Abdur Rakib and Lay-Ki Soon. Using the reddit corpus for cyberbully detection. In Ngoc Thanh Nguyen, Duong Hung Hoang, Tzung-Pei Hong, Hoang Pham, and Bogdan Trawiński, editors, *Intelligent Information and Database Systems*, pages 180–189, Cham, 2018. Springer International Publishing.

[12] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. Semi-supervised learning.

[13] CNBC. False rumor of explosion at white house causes stocks to briefly plunge; ap confirms its twitter feed was hacked. https://www.cnbc.com/id/100646197, Apr 2013.

[14] Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. Tweetcred: Real-time credibility assessment of content on twitter. In *International Conference on Social Informatics*, pages 228–243. Springer, 2014.

[15] Fang Jin, Wei Wang, Prithwish Chakraborty, Nathan Self, Feng Chen, and Naren Ramakrishnan. Tracking multiple social media for stock market event prediction. In *Industrial Conference on Data Mining*, pages 16–30. Springer, 2017.

[16] Arnav Joshi, Ravendar Lal, Tim Finin, and Anupam Joshi. Extracting cybersecurity related linked data from text. In *Proceedings of the 7th IEEE International Conference on Semantic Computing*. IEEE Computer Society Press, September 2013.

[17] Michael Kandefer, S Shapiro, Adam Stotz, and Moises Sudit. Symbolic reasoning in the cyber security domain. 2007.

[18] Ravendar Lal. Information Extraction of Security related entities and concepts from unstructured text. Master's thesis, University of Maryland, Baltimore County, May 2013.

[19] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

[20] Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. DBpedia spotlight: shedding light on the web of documents. In *7th Int. Conf. on Semantic Systems*, pages 1–8. ACM, 2011.

[21] Sudip Mittal, Prajit Kumar Das, Varish Mulwad, Anupam Joshi, and Tim Finin. Cybertwitter: Using twitter to generate alerts for cybersecurity threats and vulnerabilities. In *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*, pages 860–867. IEEE, 2016.

[22] Sudip Mittal, Anupam Joshi, and Tim Finin. Thinking, fast and slow: Combining vector spaces and knowledge graphs. *corpus*, 2:3, 2017.

[23] The Register. Most vulnerabilities first blabbed about online or on the dark web. https://www.theregister.co.uk/2017/06/08/vuln_disclosure_lag/, Jun 2017.

[24] The Register. Make america late again: Us 'lags' china in it security bug reporting. https://www.theregister.co.uk/2017/10/20/us_china_vuln_reporting/, Oct 2017.

[25] Robert D. Steele. The importance of open source intelligence to the military. *International Journal of Intelligence and CounterIntelligence*, 8(4):457–470, 1995.

[26] Rock Stevens, Octavian Suciu, Andrew Ruef, Sanghyun Hong, Michael W. Hicks, and Tudor Dumitras. Summoning demons: The pursuit of exploitable bugs in machine learning. *CoRR*, abs/1701.04739, 2016.

[27] Zareen Syed, Ankur Padia, M. Lisa Mathews, Tim Finin, and Anupam Joshi. UCO: A unified cybersecurity ontology. In *AAAI Workshop on Artificial Intelligence for Cyber Security*, pages 14–21. AAAI Press, 2015.

[28] Takeshi Takahashi, Hiroyuki Fujiwara, and Youki Kadobayashi. Building ontology of cybersecurity operational information. In *6th Workshop on Cyber Security and Information Intelligence Research*, page 79. ACM, 2010.

[29] Takeshi Takahashi, Youki Kadobayashi, and Hiroyuki Fujiwara. Ontological approach toward cybersecurity in cloud computing. In *3rd Int. Conf. on Security of information and networks*, pages 100–109. ACM, 2010.

[30] Jeffrey Undercofer, Anupam Joshi, and John Pinkston. Modeling Computer Attacks: An Ontology for Intrusion Detection. In *Proc. 6th Int. Symposium on Recent Advances in Intrusion Detection*. Springer, September 2003.

[31] Tim Weninger, Xihao Avi Zhu, and Jiawei Han. An exploration of discussion threads in social news sites: A case study of the reddit community. In *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*, pages 579–583. IEEE, 2013.