APPROVAL SHEET

Title of Thesis: Preventing Poisoning Attacks on AI based Threat Intelligence Systems

Name of Candidate: Nitika Khurana
Master of Science, 2018

Thesis and Abstract Approved: ________________________
Dr. Anupam Joshi
Oros Family Professor and Chair
Department of Computer Science and
Electrical Engineering

Date Approved: ______________________________

ABSTRACT

Title of Thesis: Preventing Poisoning Attacks on AI based Threat Intelligent Systems

Nitika Khurana, Master of Science, 2018

Thesis directed by: Dr. Anupam Joshi, Oros Family Professor
and Chair
Department of Computer Science and
Electrical Engineering

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 determine the credibility of Reddit posts by estimating their reputation score to ensure the validity of information ingested by AI systems. We also maintain the provenance of the output generated to ensure information and source reliability and identify the background data that caused an attack. 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.

Preventing Poisoning Attacks on AI based Threat Intelligent Systems

by

Nitika Khurana

Thesis submitted to the Faculty of the Graduate School
of the University of Maryland in partial fulfillment
of the requirements for the degree of
Master of Science
2018

I dedicate this work to my mom, dad, and brother.

ACKNOWLEDGMENTS

I would like to first express my deepest gratitude to my thesis advisor, Dr. Anupam Joshi for supporting me through my masters study and research. I am gratefully indebted to his invaluable guidance, understanding, patience and motivation for this thesis. I am also grateful to Sudip Mittal, who acted as my mentor throughout my thesis and provided useful insights. I would like to thank my roommate and friend, Srishty Saha, who encouraged and advised me through it. I would also like to thank my parents Rajeev Khurana and Meenu Khurana, my brother, Nikhil for being my strength. This accomplishment would not have been possible without them.

Thank You!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Artificial Intelligence 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. *Overt* sources such as newspapers, blogs, dark web, online social media (OSM), technical reports,journals, etc are consumed by AI for training. These AI systems are widely used in major industries like finance (Jin *et al.* 2017) and cybersecurity (Mittal *et al.* 2016). In stock market, AI systems are used for Algorithm trading which extracts information from OSM to execute large commands based on a pre-programmed automated trading instructions (Lin 2013).

In cybersecurity, information is mined from 'Open-source Intelligence' (OSINT) (Steele 1995). 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 for threat intelligence. Some of the most commonly used OSM are Twitter, Reddit[1], etc.

[1] https://www.twitter.com, https://www.reddit.com

Threat intelligence or cyber threat intelligence (CTI) contains an organized and refined information on potential attacks that threaten an organization. This information about potential attacks helps organizations come up with policies to prevent these attacks which are relevant to their businesses.

A new class of threat intelligence systems are being developed that use AI to extract threat intelligence. These are termed as 'Augmented Intelligence' systems. Watson for cybersecurity [2] is the first AI based security intelligence system that helps analysts identify threats more accurately and faster than ever before. These cognitive intelligence systems help security analysts identify new vulnerabilities, analyze network and endpoint activity, find evidence of preplanned attacks and hints of data breaches. They mine information from traditional OSINT sources like NIST's National Vulnerability Database (NVD)[3], United States Computer Emergency Readiness Team (US-CERT)[4], etc. and non-traditional sources like Twitter, Reddit, blogas and news.Non-traditional sources are faster than the traditional ones. There is a significant gap between initial vulnerability announcement and NVD release (Register Oct 2017). Vulnerability threat intelligence appears first on non-traditional sources (Register Jun 2017). Mining non-traditional sources is becoming really important. In our previous work, we have developed *CyberTwitter* (Mittal *et al.* 2016) that mines threat intelligence from Twitter.

The very 'open' nature of these OSINT sources is two-edged. They are vulnerable to misinformation in the form of hoaxes, false images and videos, and rumors. This traditionally constitutes as fake news (Lazer *et al.* 2018). This leaves the organizations susceptible

[2] https://www.ibm.com/security/cognitive

[3] https://nvd.nist.gov/

[4] https://www.us-cert.gov/

to ‘poisoning attacks’ by a malicious entity. Attackers can ‘poison’ the data used for intelligence by adding incorrect information to get past their cyber defenses. In a recent poisoning attack on Twitter, a tweet was posted by Associated Press claiming that US’s then President, Barak Obama had been injured in a series of bomb blasts at the White House. Algorithm trading systems read this tweet and started selling S&P futures and buying Treasury 10-year futures. 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 (CNBC Apr 2013). This traditionally constitutes as fake news (Lazer *et al.* 2018). Several of these fake news incidents have caused a loss of money, reputation, infrastructure and in certain cases, threat to human lives.

In another incident, a hacker hacked into the Qatari news sites and social medias and

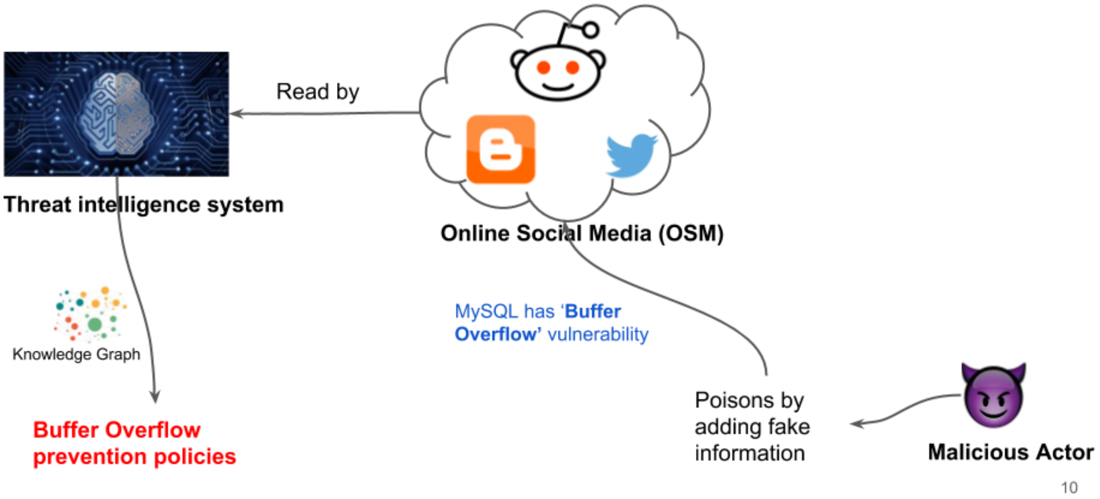

Read by
Threat intelligence system
Online Social Media (OSM)
Knowledge Graph
MySQL has ‘Buffer Overflow’ vulnerability
Buffer Overflow prevention policies
Poisons by adding fake information
Malicious Actor
10

FIG. 1.1. Attacker model: adding fake information.

posted false comments attributed to Qatars emir, Tamim bin Hamad al-Thani. The posts cited him criticizing Donald Trump and praising Iran with the intention of damaging the image and reputation of Qatar (Quartz Oct 2017). This led to Qatar's 5 neighbouring states to break their relations with Qatar and block truck, ship and air traffic into the country (Windrem & William M. Arkin Jul 2017).

Increasing adoption of these non-traditional sources in AI cyber defense systems have created a potential attack surface. Attackers may employ two models to 'poison' the data: by adding fake and contradictory information. For example, an attacker might spread the information like there exists a buffer overflow in Mozilla Firefox, this fake information might trigger a policy change directive by the defensive AI. An attacker might use this as a diversionary tactic against the AI.

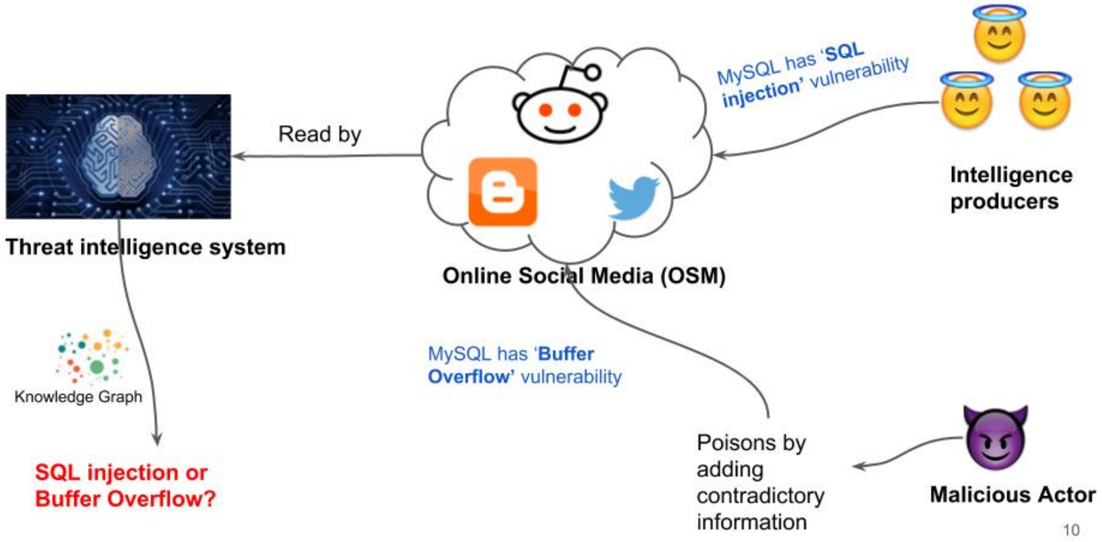

MySQL has 'SQL injection' vulnerability
Read by
Threat intelligence system
Online Social Media (OSM)
Intelligence producers
Knowledge Graph
MySQL has 'Buffer Overflow' vulnerability
SQL injection or Buffer Overflow?
Poisons by adding contradictory information
Malicious Actor
10

FIG. 1.2. Attacker model: adding contradictory information.

In another model, they can also put in contradicting information about a valid threat intelligence. for example, an attacker might put the information out that a buffer overflow vulnerability exists in software MySQL, wherein software MySQL has a SQL injection vulnerability. In this case, the contradicting information will harm the AI system making it more susceptible to attack. Figure 1 and 1 explains the above two attacker models.

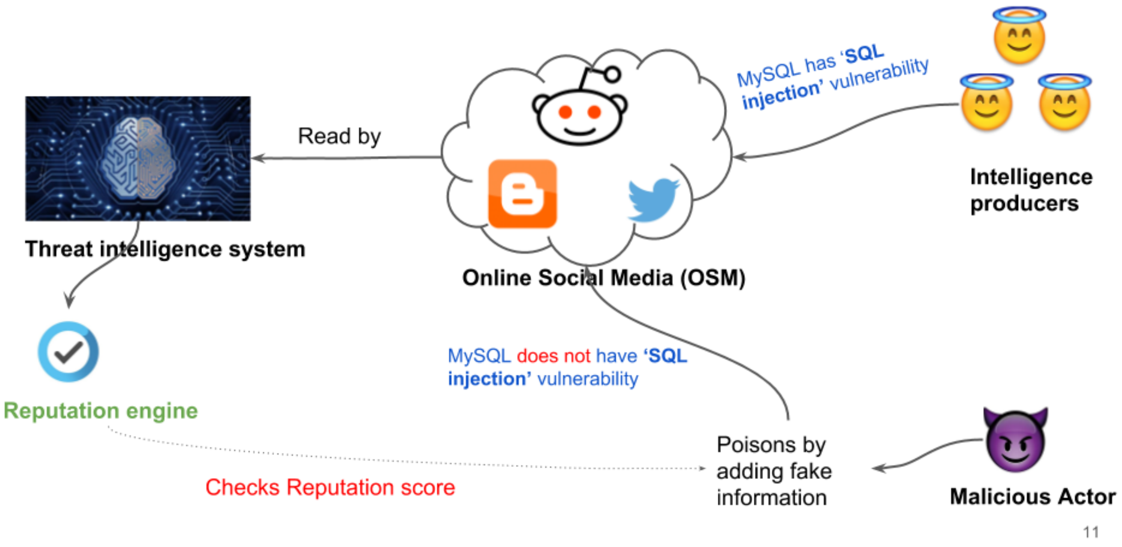

MySQL has 'SQL injection' vulnerability
Read by
Intelligence producers
Threat intelligence system
Online Social Media (OSM)
MySQL does not have 'SQL injection' vulnerability
Reputation engine
Poisons by adding fake information
Malicious Actor
Checks Reputation score
11

FIG. 1.3. Attack scenario and proposed defense. Fake or contradictory information added by Attacker is verified using a reputation engine.

This information, if consumed by the AI cyber defense system, can help the attacker evade various security measures putting the organization at risk. Figure 1 explains this attack scenario and proposed defense. In this research, we propose 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. More details about our proposed engine are described in Section 3.

Another way to maintain trustworthiness of the gathered data is to encode its provenance in the knowledge representation. Provenance data can help a system analyst to identify the exact data source, the time when the data was last updated, etc. This provenance data will be a useful tool for the security analyst, when she tries to investigate recommendations and changes suggested by the cyber-defense AI. It gives analyst the trigger i.e. she can interfere and suggest AI based Cyber Threat Intelligence (CTI) against the change. In this paper, we employ a RDF based approach for maintaining provenance.

The remaining document is organized as follows: Section 2 describes the background and the related work. Section 3 discusses our methodology for data collection, annotation scheme used for establishing the ground truth, feature selection and reputation score calculation. Section 4 summarizes our results. We conclude in Section 5.

Chapter 2

RELATED WORK & BACKGROUND

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

2.1 Background

In this section, we describe Online Social Media (OSM), prevalent fake information on OSM, impact of incredible data in cybersecurity, threat intelligence systems, Reddit platform and significance of provenance data.

2.1.1 Online Social Media (OSM)

Online social Media (OSM) provides a platform for users to disseminate their opinion on topics like news, sports, entertainment, earthquakes, politics, art, culture, etc. , build personal and professional networks and advertise or share information. There are several existing OSM - Facebook[1], is the most commonly used social platform to connect with friends by the users, Twitter[2] is a social news and networking site to send messages (or most commonly called ”tweets”), LinkedIn[3] is a professional networking platform to provide or

[1] https://www.facebook.com

[2] https://www.twitter.com

[3] https://www.linkedin.com

consume available employment opportunities, YouTube[4] is a video sharing platform and Reddit[5] remains a social news aggregation website. The widespread use of OSM has made it possible to transmit or share data in milliseconds of time. OSM primarily deals with 4Vs of data:

- Velocity: Velocity denotes the analysis of streaming data. The huge streaming data needs to be analyzed for malicious activity, spam and phishing data.
- Volume: Volume refers to the scale of data. With the increase in the number of users consuming the services of OSM, the sheer amount of data being produced is huge. A huge amount of content is generated every second, minute and hour of the day and hence, need a scalable system for analysis.
- Variety: Variety defines the different forms of data. The huge amount of OSM data is available in structured as well as unstructured form. Publicly available overt sources include newspapers, magazines, social networking sites, video sharing sites, wikipedia, blogs, etc.
- Veracity: Veracity denotes the uncertainty of data. With the huge amount of data being produced, it is difficult to analyze the trustworthiness of it. Internet is an anonymous community and hence anyone can create a profile on OSM without verification.

2.1.2 Reddit

Reddit is a social news aggregation, web content rating, and discussion website with over 230 million users (Weninger, Zhu, & Han 2013). On Reddit, registered users can post text, URLs, and images to which other user users can like, comment or

[4] `https://www.youtube.com`

[5] `https://www.reddit.com`

down-vote. Reddit organizes the posts into user-created boards called ‘subreddits’; addressing varying topics like news, science, security, movies, video games, gadgets, music, education, books, fitness, food, etc. The data is segregated into different tabs within each ‘subreddit’ and if a post received enough up-votes, it can be seen on the site’s front page. The members of Reddit commuity are also called as ‘Redditors’. Redditors earn points for submitting comments, links, text posts called as “comment karma”, “link karma” and “post karma” respectively.

2.1.3 AI based Threat Intelligence Systems

Threat intelligence or cyber threat intelligence (CTI), is organized, analyzed and refined information about potential or current attacks that threaten an organization. Provide organizations with current information related to potential attack sources relevant to their businesses; some also offer consultation service. Because of the huge size of real-time data available, it is impossible to organize and analyze it manually or via ad-hoc systems and thus, security analysts use Artificial Intelligence based organizational cyber-defense systems, also termed as “Augmented Intelligence Systems” (by IBM 2018). These systems are used by security analysts to assimilate, correlate, and analyze potential threats or cyber attacks from varied information sources. These systems utilize real-time data to identify potential risks relevant to their organization to devise defensive and corrective measures.

2.1.4 Word Embedding

Word embeddings are used to represent words in a continuous vector space. Two popular methods to generate these embeddings are word2vec (Nickel, Rosasco, & Poggio 2015), (Mikolov *et al.* 2013) and GloVe (Pennington, Socher, & Manning

2014). The main idea behind generating embeddings for words is to say that vectors close together are semantically related. Word embeddings have been used in various applications like machine translation, improving local and global context, etc.

2.2 Related Work

2.2.1 AI for Cybersecurity

Knowledge graphs have been used in cybersecurity to combine data and information from multiple sources, these systems then aid a security analyst in her day to day operations. Various ontology based intrusion detection systems (Undercofer, Joshi, & Pinkston 2003, Kandefer *et al.* 2007, Takahashi, Kadobayashi, & Fujiwara 2010, Takahashi, Fujiwara, & Kadobayashi 2010) have been put forth by researchers. These systems depend on a data repository of system vulnerabilities and threats (Joshi *et al.* 2013, Mittal *et al.* 2016). These repositories are stored as RDF[6] linked data created from vulnerability descriptions collected from the National Vulnerability Database, Twitter, etc. Joshi et al. (Joshi *et al.* 2013) 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 (Auer *et al.* 2007) using DBpedia Spotlight (Mendes *et al.* 2011). CyberTwitter (Mittal *et al.* 2016), a framework to automatically issue cybersecurity vulnerability alerts to users. CyberTwitter converts vulnerability intelligence from tweets to RDF. It uses the UCO ontology (Unified Cybersecurity Ontology) (Syed *et al.* 2015) to provide their system with cybersecurity domain information. Mittal et al. have also created *Cyber-All-Intel* where they have used multiple knowl-

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

edge representations to store threat intelligence (Mittal, Joshi, & Finin 2017).

Systems like the one proposed in (Mittal *et al.* 2016, Mittal, Joshi, & Finin 2017) 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. For example, 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 3.

2.2.2 Attacks on AI

AI systems are susceptible to threats posed by malicious inputs (Register Jun 2017), (Register Oct 2017). Stevens et al. (Stevens *et al.* 2016) describes how malicious inputs exploiting implementation bugs in ML algorithms poses 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.

2.2.3 Credibility of Threat 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’ (Gupta *et al.* 2014), 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’ particularly the 2010 earthquake in Chile (Mendoza, Poblete, & Castillo 2010), the Hurricane Sandy in 2012 (Gupta *et al.* 2013) and the Boston Marathon blasts in 2013 (Gupta, Lamba, & Kumaraguru 2013) and thereby adversely affecting thousands of people. The model used for TweetCred is a semi-supervised ranking model that uses SVM-rank to identify the credibility of data based on 45 identified features. Rakib et al. used word embeddings on Reddit database based on word2vec skip-gram model to train a random forest classifier to identify cyberbully comments (Bin Abdur Rakib & Soon 2018). 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 credibility score (See Section 3).

2.2.4 Provenance

Provenance data can be beneficial to identify the steps or the ‘background data’ that caused an attack. Moonesinghe et al. (Moonesinghe, Khoury, & Janssens 2007) showed that reproducibility of data can benefit in improving the quality of research. To address this clause, a PROV (Moreau *et al.* 2013) tool was developed by W3C to track the provenance of artifacts. The PROV data model is a conceptual data model that defines the provenance specifications for PROV. It consists of six components dealing with entities or events with their timing life-cycles, derived entities, provenance of provenance, entities referring to same thing, etc. This model can be mapped to RDF using OWL 2 (Web Ontology Language) (De Nies *et al.* 2013). This representation will be used to represent provenance trees for our system. For reproducibility of big data experiments, PROB tool was devised by Korolev et al. (Korolev, Joshi, & Grasso 2014) which integrates Git2Prov, Git and Git-Annex (Hess accessed 04 January 2014) and defines its own ontology for provenance representation. Ding Li et al. (Ding *et al.* 2005) disintegrated provenance information represented using RDF graph into RDF molecules.

Provenance trees can be integrated with Proof Markup Language (PML) (Da Silva, McGuinness, & Fikes 2006) (now called as Provenance Markup Language) to define a provenance ontology that defines the representational primitives to define the attributes of information, language, and sources such as a person, an organization, text, etc. McGuinnes et. al (Da Silva *et al.* 2008) described three additional vocabularies for PML to include provenance data (PML-P), justification data (PML-J) and a trust relation ontology (PML-T) and named the extended PML as PML 2. The provenance ontology of PML 2 will provide the necessary information about the origin of data and thus, help in user understanding of generated outputs and will facilitate user acceptance of the outputs.

Chapter 3

METHODOLOGY

In this section, we describe the overall architecture (See Figure 3.1) of our proposed system that includes a reputation engine to calculate the credibility score for each post. The system was created by generating a set of features to train our model on a manually annotated ground truth training set. We use a supervised learning algorithm. The reputation score is generated using the distance of a post's embeddings from 'credible' and 'non-credible' clusters.

3.1 Data Collection

We collected data from Reddit using the PRAW[1] API which is a python Reddit API Wrapper. PRAW 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 4500 posts over a span of last two months corresponding to several cybersecurity subreddits: cybersecurity, malware, cryptography, cryptocurrency, cyber, cryptomarkets, cyberlaw and cybersecurityfans.

[1] https://praw.readthedocs.io/en/latest/index.html

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

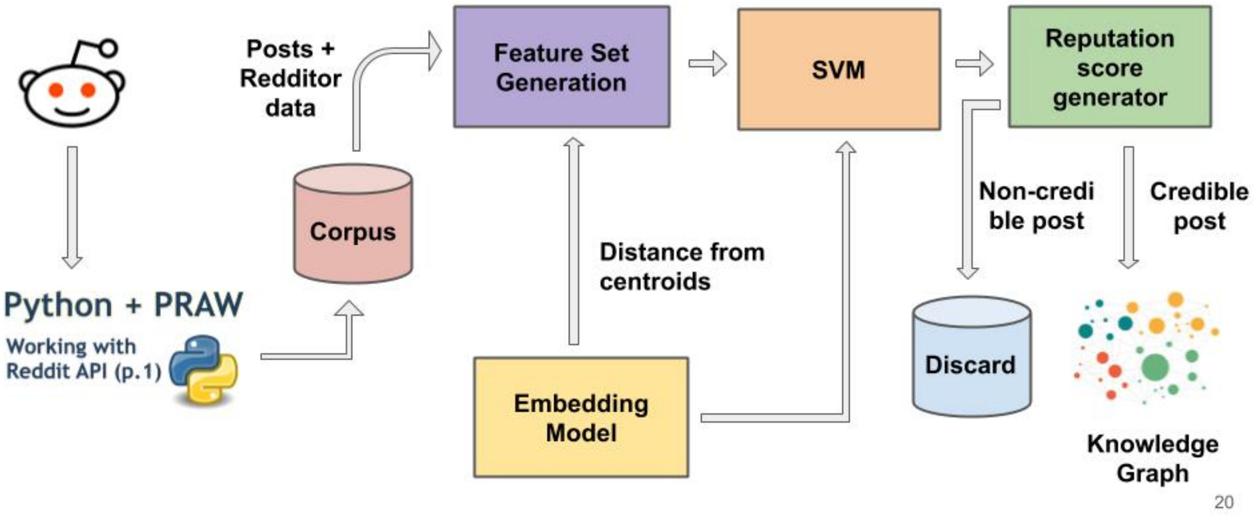


FIG. 3.1. Architecture of our methodology and analysis.

3.2 Annotation & Feature Selection

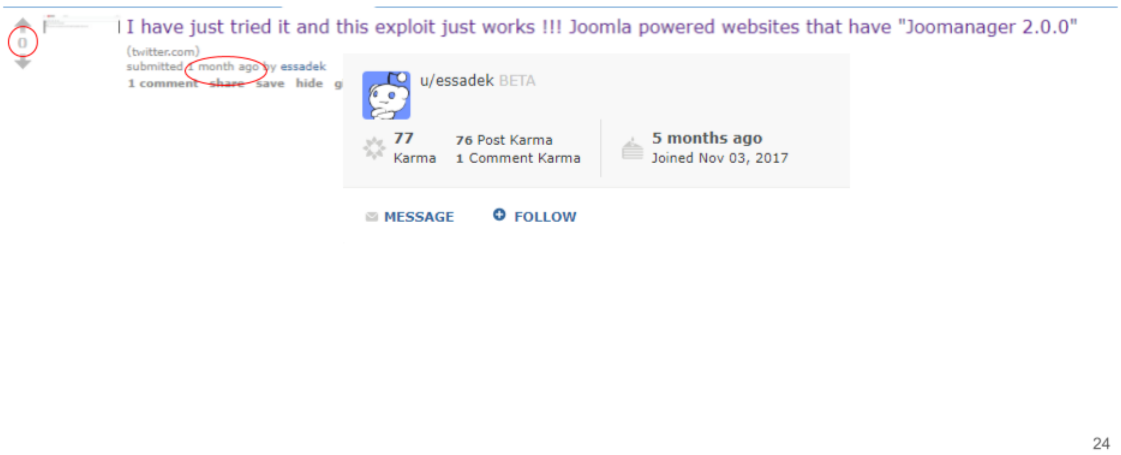
Human annotators were used to obtain the ground truth for our experiments. Human annotation is a research methodology well-known for establishing the ground truth (Krig 2014). From the 4500 posts collected over a span of last two months, we randomly picked a sample of 2000 posts for annotation. We provided the annotators the definition of credibility and asked them to classify the posts into two classes: ‘credible’ or ‘non-credible’. A ‘credible’ post is one that contains true information. Annotators were given added information like referred Common Vulnerabilities and Exposures (CVE) database entries and links to verified news websites like The Washington Post (was), BBC (bbc), The Guardian (gua), CNN (cnn), Reuters (reu), etc. or cybersecurity sources like HackerNews (hac), Krebs on Security (kre), Microsoft

Feature set	**Features**
Post Features	Post Length, post time, downvotes, upvotes, downvotes & karma score, number of comments, number of crossposts (to another subreddit) and Web of Trust (WOT)[2] values of URLs
Redditor Features	‘Redditorś’ screen name length, user registration time, link karma, comment, verified user email, verified user, user is a moderator or not (responsible for organizing all posts in a specific subreddit).

Table 3.1. Selected features for analysis.

(mic), etc. We selected a set of 15 features corresponding to the post and ‘Redditor’. Table 3.1 lists the features accessed via the PRAW API for all posts and the ‘Redditors’. Also, the distance of the post vectors from the centroid of the two clusters of ground truth was used as another feature for our classifier.

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. 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 around 1206 posts that served as ground truth with 953 posts entitled as ‘credible’ and 253 as ‘non-credible’. These results were generally based on three attributes of a post. The annotators were asked to pen down their criteria for establishing the credibility of each post. On analyzing their data, we found that the annotators weighted the credibility of the URL 50% of the time, with the post’s content verified via credible sources was considered 35% of the time and the Redditor’s features were evaluated for the remaining posts. Vectors generated in Section 3.3 have also been included in our feature set.



I have just tried it and this exploit just works !!! Joomla powered websites that have "Joomanager 2.0.0"

(twitter.com)

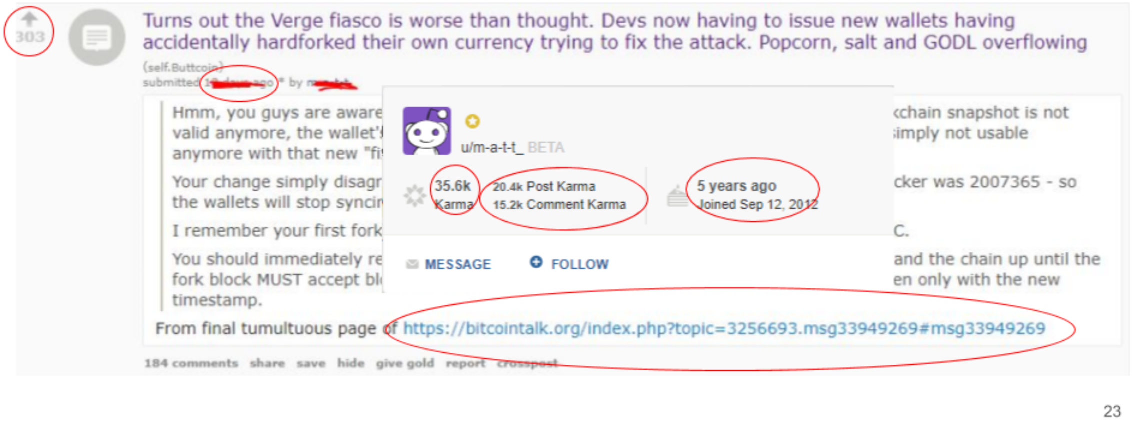
submitted 1 month ago by essadek

FIG. 3.2. A ‘non-credible’ Reddit post in ground truth. The post has 0 likes, non-descriptive with no URL and the ‘Redditor’ has no karma.

3.3 Vector generation

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. We first used an NER to identify cybersecurity terms. The word embeddings were taken from the model created by Mittal et al. for their *Cyber-All-Intel* system (Mittal, Joshi, & Finin 2017).

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 4.1.


303

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

(self.Buttcoin)
submitted [redacted]* by [redacted]

Hmm, you guys are aware
valid anymore, the wallet'
anymore with that new "fi

Your change simply disagr
the wallets will stop syncin

I remember your first fork

You should immediately re
fork block MUST accept bl
timestamp.

u/m-a-t-t_ BETA

35.6k Karma | 20.4k Post Karma 15.2k Comment Karma | 5 years ago Joined Sep 12, 2012

kchain snapshot is not
simply not usable

cker was 2007365 - so

C.

and the chain up until the
en only with the new

From final tumultuous page of https://bitcointalk.org/index.php?topic=3256693.msg33949269#msg33949269

FIG. 3.3. A ‘Credible’ Reddit post observed by manual annotation. The post has many likes, high WOT score and high link and comment karma of the ‘Redditor’.

3.4 Reputation score generation

In our system, we wish to create a quantifiable score which can be understood by both the AI system and the security analyst. We begin by defining the feature set (Section 3.2 & 3.3) and then train a classifier using Linear Support Vector Machine (SVM) to determine the credibility of a post. After training the model, we classified the posts into two classes ‘credible’ and ‘non-credible’. We then calculate the reputation score of a post by determining the distance of the post vector from the cluster centroids created in Section 3.3. 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_{nc}) as:

$$s_c = 1 - \frac{d_c}{d_c + d_{nc}}$$

We use both the 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 4.

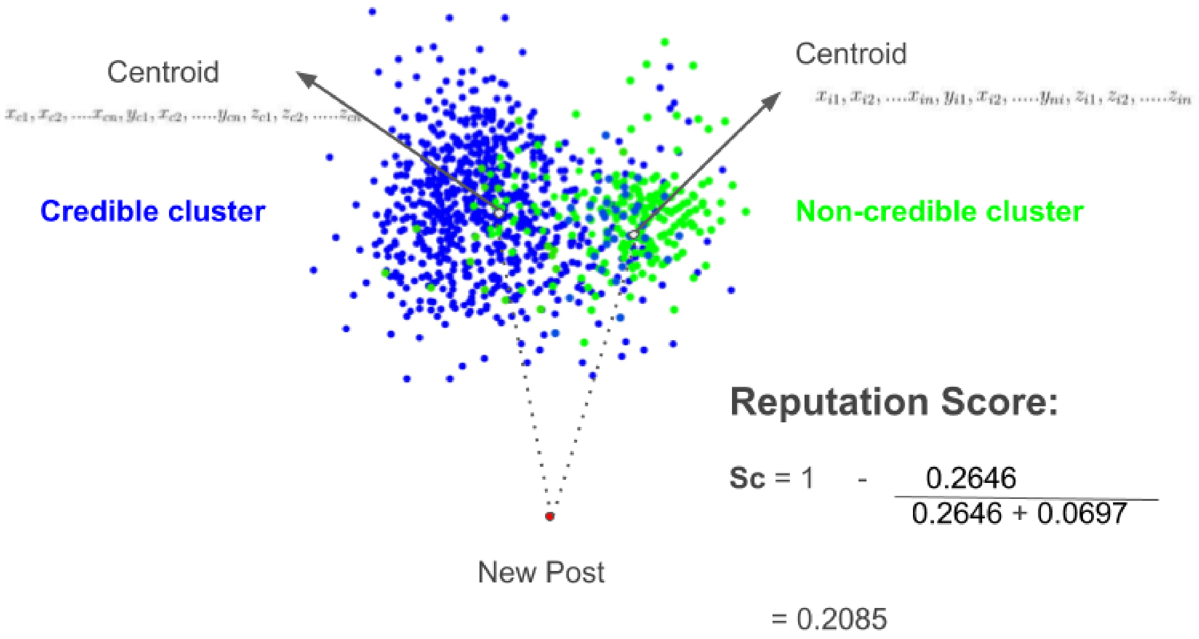

Centroid
$x_{c1}, x_{c2}, ...x_{cn}, y_{c1}, x_{c2}, ...y_{cn}, z_{c1}, z_{c2}, ...z_{cn}$
Centroid
$x_{i1}, x_{i2}, ...x_{in}, y_{i1}, x_{i2}, ...y_{ni}, z_{i1}, z_{i2}, ...z_{in}$
Credible cluster
Non-credible cluster
Reputation Score:
$\mathbf{Sc} = 1 - \frac{0.2646}{0.2646 + 0.0697}$
$= 0.2085$
New Post

FIG. 3.4. Reputation score generation for a new post.

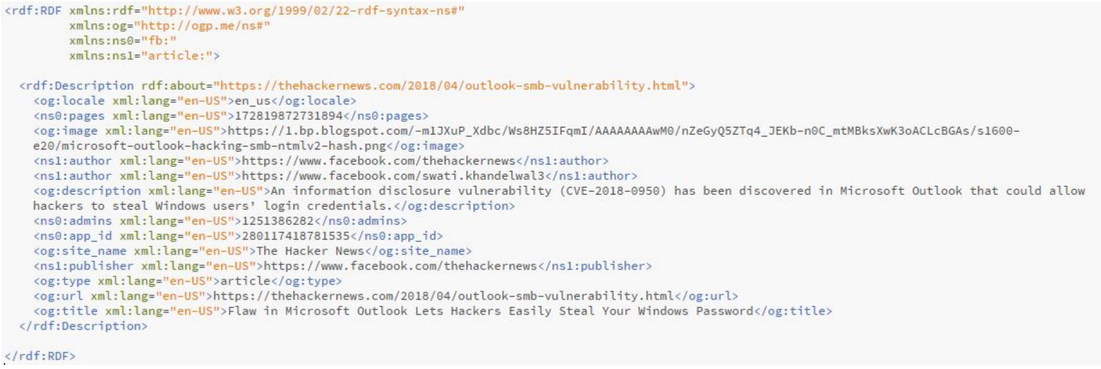
The algorithm 3.4 describes all the above methods of calculating the reputation score of posts. Function $ExtractFeatures(R)$ computes the content and ‘Redditor’ features for each post r_i in a set of posts R. The $WordEmbedding(R)$ function evaluates the word embeddings for each post using the embedding model similar to the one defined by Cyber-All-Intel (Mittal, Joshi, & Finin 2017). The

Algorithm 1 ReputScore ($R[1..n], A[1..m]$)

```
Require: $Centroids \leftarrow VectorClusters(A[1..m])$
  for $i = 1$ to $n - 1$ do
    $F_i \leftarrow ExtractFeatures(R[\text{i}])$
  end for
  for $i = 1$ to $n - 1$ do
    $W_i \leftarrow WordEmbedding(R[\text{i}])$
  end for
  $Cred \leftarrow LinearSVM(F)$
  $Score \leftarrow EvalDistance(SUM(W), Centroids)$
  return $Score$
```

word embedding projections are then added together to get the vector for the entire post. $VectorClusters(A)$ takes the ground truth post's vectors to generate two clusters and outputs their centroids (see section 3.3). $LinearSVM(F)$ takes the feature set for the posts and classifies posts as ‘credible’ or ‘non-credible’. $EvalDistance(W, Centroids)$ function measures the distance between the cluster centroids and a new post vector to determine its reputation score.

3.5 Provenance information generation



```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:og="http://ogp.me/ns#"
         xmlns:ns0="fb:"
         xmlns:ns1="article:">

  <rdf:Description rdf:about="https://thehackernews.com/2018/04/outlook-smb-vulnerability.html">
    <og:locale xml:lang="en-US">en_us</og:locale>
    <ns0:pages xml:lang="en-US">172819872731894</ns0:pages>
    <og:image xml:lang="en-US">https://1.bp.blogspot.com/-m1JXuP_Xdbc/Ws8HZ5IFqmI/AAAAAAAAwM0/nZeGyQ5ZTq4_JEKb-n0C_mtMBksXwK3oACLcBGAs/s1600-
    e20/microsoft-outlook-hacking-smb-ntmlv2-hash.png</og:image>
    <ns1:author xml:lang="en-US">https://www.facebook.com/thehackernews</ns1:author>
    <ns1:author xml:lang="en-US">https://www.facebook.com/swati.khandelwal3</ns1:author>
    <og:description xml:lang="en-US">An information disclosure vulnerability (CVE-2018-0950) has been discovered in Microsoft Outlook that could allow
    hackers to steal Windows users’ login credentials.</og:description>
    <ns0:admins xml:lang="en-US">1251386282</ns0:admins>
    <ns0:app_id xml:lang="en-US">280117418781535</ns0:app_id>
    <og:site_name xml:lang="en-US">The Hacker News</og:site_name>
    <ns1:publisher xml:lang="en-US">https://www.facebook.com/thehackernews</ns1:publisher>
    <og:type xml:lang="en-US">article</og:type>
    <og:url xml:lang="en-US">https://thehackernews.com/2018/04/outlook-smb-vulnerability.html</og:url>
    <og:title xml:lang="en-US">Flaw in Microsoft Outlook Lets Hackers Easily Steal Your Windows Password</og:title>
  </rdf:Description>

</rdf:RDF>
```

FIG. 3.5. Provenance: RDF instance of a Reddit post.

To ensure that the system analyst is able to understand the recommendations generated by an AI, we add provenance to the threat intelligence RDF. For example, the AI may recommend a policy update to the analyst, who may want to investigate deeper as to why and what intelligence precipitated the change directive.

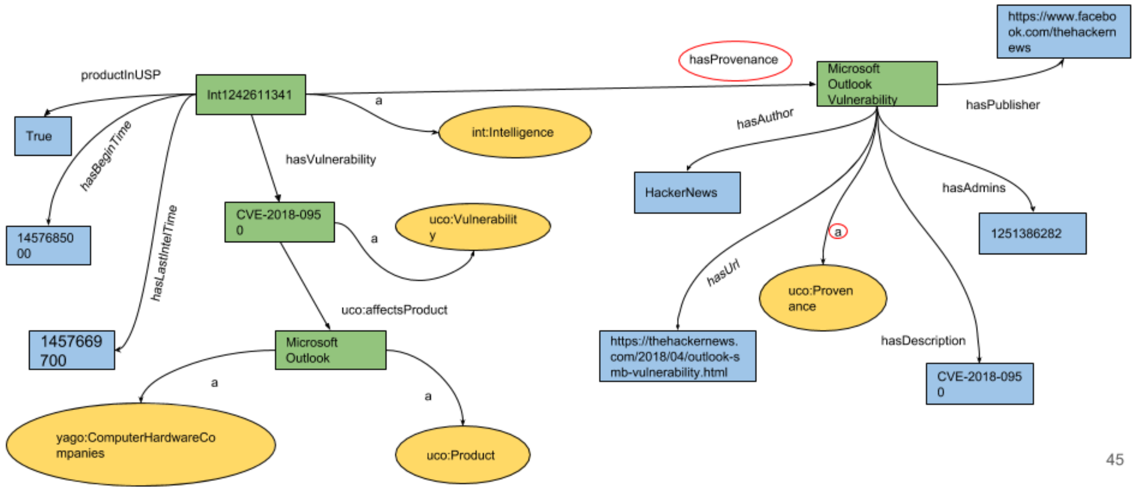


FIG. 3.6. A new class ‘uco:Provenance’ is added to Unified Cybersecurity Ontology (UCO). Provenance RDF is linked to Threat Intelligence RDF using the ‘hasProvenance’ property.

To generate the provenance of the data, we create RDF statements for the provenance data present in Reddit posts. Figure 3.5 shows details of the provenance for a Reddit post. The post’s RDF instance includes attributes such as the post’s author, description, URL’s publisher and description, etc. This graph is linked to the intelligence RDF using ‘hasProvenance’ property as an RDF molecule. So, in case of a policy change by an AI, a security analyst utilizes this property to access the provenance graph describing the attributes that led to this policy change. Figure 3.6 shows how provenance is linked using the property ‘hasProvenance’ to the threat intelligence RDF.

Chapter 4

RESULTS

This section describes the results obtained on classifying posts using Linear SVM. We explain the accuracy of our model and the features that turned out to be strong indicators of credibility.

4.1 Classification Analysis

We performed Linear Support Vector Machine (SVM) over the selected features described in Table 3.1 to estimate the credibility of the posts. After training Linear SVC on the annotated 1206 posts, we obtained a learned model that classifies posts for credibility.

We, then, evaluated ten-fold cross validation of our results. The dataset is partitioned into 9 different sets of training data with a single subsample of the data used for validation. Over a training set of 1206 posts, the 10 results from the folds was averaged (or combined) to give us an accuracy of 87.73%.

Table 4.1 describes the confusion matrix obtained for the predicted posts. Out of 953 credible posts, we correctly identified 851 to be credible and 96 turned out to

	Positive	**Negative**
Predicted Positive	851	96
Predicted Negative	52	207

Table 4.1. Confusion matrix for a set of 80:20 training and test data.

Derived metrics	**Values**
Accuracy	87.728%
Precision	0.68317
Recall	0.79923
Error Rate	12.272%
True Negative Rate	89.863%
False Positive Rate	10.137%
F1 Score	0.73665
F0.5 Score	0.70360
F2 Score	0.77296

Table 4.2. Confusion matrix and derived metrics for 80:20 training and test data.

be falsely predicted as credible. Also, 207 were correctly identified as incredible out of the 253 negative posts. Thus, our analysis for credibility predicted results with an accuracy of 87.73%. Table 4.2 shows the derived metrics from confusion matrix and their values. We also computed the confusion matrix as shown in Table 4.3 and its derived metrics (Table 4.4) for a balanced set of ‘credible’ and ‘non-credible’ posts. The comparatively greater number of false negatives than false positives justifies our methodology. This implies that there are more number of ‘credible’ posts classified as ‘non-credible’ than ‘credible’ thereby preventing our system from potential poisoning attacks.

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

	Positive	**Negative**
Predicted Positive	188	67
Predicted Negative	77	174

Table 4.3. Confusion matrix for balanced set of 253 ‘credible’ and 253 ‘non-credible’ posts.

Derived metrics	**Values**
Accuracy	71.541%
Precision	0.72199
Recall	0.69323
Error Rate	28.458%
True Negative Rate	73.725%
False Positive Rate	26.274%
F1 Score	0.70732
F0.5 Score	0.71605
F2 Score	0.69879

Table 4.4. Confusion matrix and derived metrics for a balanced set.

from which the data is extracted. High WOT score websites are observed to be the verified news websites like The Washington Post (was), BBC (bbc), The Guardian (gua), CNN (cnn), Reuters (reu), etc. or cybersecurity sources like HackerNews (hac), Krebs on Security (kre), Microsoft (mic), 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.

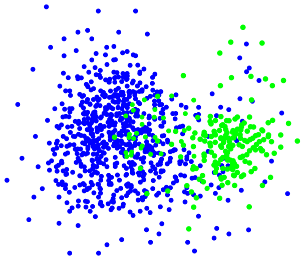


FIG. 4.1. Visualization of post clusters using t-SNE. Blue cluster represents 'credible' posts and green represents 'non-credible' annotated posts.

We also calculated the reputation score of the posts using their relative distances from the credible and incredible clusters obtained from ground truth post vectors. Figure 4.1 shows that posts identified as 'credible' by classification tend to lie in close proximity of credible cluster (colored red) and 'non-credible' posts lie close to incredible cluster (colored blue). The distance from the centroids of the two clusters for two sample posts is listed in Table 4.5. The post titled 'I have just tried it and this exploit just works !!! Joomla powered websites that have "Joomanager 2.0.0"' was identified as 'non-credible' by our analysis and was closer to the non-credible cluster and the second post was closer to the credible cluster and identified as '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 (calculated using equation) of 0.2085 and post 2 received a score of 0.8969.

Title	**Distance from Credible cluster (units)**	**Distance from Non-credible cluster (units)**	**Reputation score**
I have just tried it and this exploit just works !!! Joomla powered websites that have “Joomanager 2.0.0”	0.2646	0.0697	0.2085
Turns out the Verge fiasco is worse than thought. Devs now having to issue new wallets having accidentally hard-forked their own currency trying to fix the attack. Popcorn, salt and GODL overflowing	0.0343	0.2986	0.8969

Table 4.5. Distance of post’s vector from centroid of two clusters.

Chapter 5

CONCLUSION & 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 and cyberlaw subreddits. We extracted Reddit posts and identified a feature set of 16 features that were trained using Linear SVM. Ground truth was established using manual annotation of around 1200 posts that were used to train our model and predict the credibility of posts. We classified the posts as ‘credible’ or ‘incredible’ with an accuracy of over 87%. 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.

We also maintain post provenance information that can be used by a security analyst to understand various policy updates and suggestions. We include provenance information by adding the provenance RDF using the ‘hasProvenance’ property. This links the system’s ontology to the provenance graph of the posts.

In the future, we would establish more ground truth data for our analysis to further improve the accuracy of our system. Also, we would like to incorporate other online social networks like Quora (quo), Twitter, dark web, etc. as they are widely used for discussions about cybersecurity threats and vulnerabilities. To establish the credibility of the URLs, we plan to incorporate analysis from Virustotal [1] and MXToolbox [2] along with WOT scores. 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.

[1] https://www.virustotal.com/

[2] https://www.mxtoolbox.com/

REFERENCES

[1] Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; and Ives, Z. 2007. *DBpedia: A Nucleus for a Web of Open Data*. Springer.

[2] Bbc: Cybersecurity.

[3] Bin Abdur Rakib, T., and Soon, L.-K. 2018. Using the reddit corpus for cyberbully detection. In Nguyen, N. T.; Hoang, D. H.; Hong, T.-P.; Pham, H.; and Trawiński, B., eds., *Intelligent Information and Database Systems*, 180–189. Cham: Springer International Publishing.

[4] by IBM, S. I. 2018. Nearly one-third of cisos have adopted ai in response to cybersecurity news, study finds.

[5] CNBC. Apr 2013. False rumor of explosion at white house causes stocks to briefly plunge; ap confirms its twitter feed was hacked.

[6] Cnn: Cybersecurity.

[7] Da Silva, P. P.; McGuinness, D.; Del Rio, N.; and Ding, L. 2008. Inference web in action: Lightweight use of the proof markup language. In *International semantic web conference*, 847–860. Springer.

[8] Da Silva, P. P.; McGuinness, D. L.; and Fikes, R. 2006. A proof markup language for semantic web services. *Information Systems* 31(4):381–395.

[9] De Nies, T.; Magliacane, S.; Verborgh, R.; Coppens, S.; Groth, P.; Mannens, E.; and Van de Walle, R. 2013. Git2prov: exposing version control system content as w3c prov. In *Proceedings of the 2013th International Conference on Posters & Demonstrations Track-Volume 1035*, 125–128. CEUR-WS. org.

[10] Ding, L.; Finin, T.; Peng, Y.; Da Silva, P. P.; and McGuinness, D. L. 2005. Tracking rdf graph provenance using rdf molecules. In *Proc. of the 4th International Semantic Web Conference (Poster)*, 42.

[11] The guardian: Technology.

[12] Gupta, A.; Lamba, H.; Kumaraguru, P.; and Joshi, A. 2013. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd international conference on World Wide Web*, 729–736. ACM.

[13] Gupta, A.; Kumaraguru, P.; Castillo, C.; and Meier, P. 2014. Tweetcred: Real-time credibility assessment of content on twitter. In *International Conference on Social Informatics*, 228–243. Springer.

[14] Gupta, A.; Lamba, H.; and Kumaraguru, P. 2013. $1.00 per # bostonmarathon # prayforboston: Analyzing fake content on twitter. In *eCrime Researchers Summit (eCRS), 2013*, 1–12. IEEE.

[15] The hacker news.

[16] Hess, J. accessed 04-January-2014. git-annex.

[17] Jin, F.; Wang, W.; Chakraborty, P.; Self, N.; Chen, F.; and Ramakrishnan, N. 2017. Tracking multiple social media for stock market event prediction. In *Industrial Conference on Data Mining*, 16–30. Springer.

[18] Joshi, A.; Lal, R.; Finin, T.; and Joshi, A. 2013. Extracting cybersecurity related linked data from text. In *Proceedings of the 7th IEEE International Conference on Semantic Computing*. IEEE Computer Society Press.

[19] Kandefer, M.; Shapiro, S.; Stotz, A.; and Sudit, M. 2007. Symbolic reasoning in the cyber security domain.

[20] Korolev, V.; Joshi, A.; and Grasso, M. 2014. Prob: A tool for tracking provenance and reproducibility of big data experiments. In *Proceeding of Workshop on Reproducible Research Methodologies (REPRODUCE'14)*.

[21] Krebs on security.

[22] Krig, S. 2014. *Ground Truth Data, Content, Metrics, and Analysis*. Berkeley, CA: Apress. 283–311.

[23] Lazer, D. M.; Baum, M. A.; Benkler, Y.; Berinsky, A. J.; Greenhill, K. M.; Menczer, F.; Metzger, M. J.; Nyhan, B.; Pennycook, G.; Rothschild, D.; et al. 2018. The science of fake news. *Science* 359(6380):1094–1096.

[24] Lin, T. C. W. 2013. The new investor. *60 UCLA Law Review 678 (2013), Temple University Legal Studies Research Paper No. 2013-45*.

[25] Mendes, P. N.; Jakob, M.; García-Silva, A.; and Bizer, C. 2011. DBpedia spotlight: shedding light on the web of documents. In *7th Int. Conf. on Semantic Systems*, 1–8. ACM.

[26] Mendoza, M.; Poblete, B.; and Castillo, C. 2010. Twitter under crisis: Can we trust what we rt? In *Proceedings of the first workshop on social media analytics*, 71–79. ACM.

[27] Microsoft: Cybersecurity.

[28] Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

[29] Mittal, S.; Das, P. K.; Mulwad, V.; Joshi, A.; and Finin, T. 2016. 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*, 860–867. IEEE.

[30] Mittal, S.; Joshi, A.; and Finin, T. 2017. Thinking, fast and slow: Combining vector spaces and knowledge graphs. *corpus* 2:3.

[31] Moonesinghe, R.; Khoury, M. J.; and Janssens, A. C. J. 2007. Most published research findings are falsebut a little replication goes a long way. *PLoS medicine* 4(2):e28.

[32] Moreau, L.; Missier, P.; Belhajjame, K.; BFar, R.; Cheney, J.; Coppens, S.; Cresswell, S.; Gil, Y.; Groth, P.; Klyne, G.; et al. 2013. Prov-dm: The prov data model. *Retrieved July* 30:2013.

[33] Nickel, M.; Rosasco, L.; and Poggio, T. A. 2015. Holographic embeddings of knowledge graphs. *CoRR* abs/1510.04935.

[34] Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation.

[35] Quartz. Oct 2017. The fake-news hack that nearly started a war this summer was designed for one man: Donald trump.

[36] Quora.

[37] Register, T. Jun 2017. Most vulnerabilities first blabbed about online or on the dark web.

[38] Register, T. Oct 2017. Make america late again: Us ’lags’ china in it security bug reporting.

[39] Reuters: Cybersecurity.

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

[41] Stevens, R.; Suciu, O.; Ruef, A.; Hong, S.; Hicks, M. W.; and Dumitras, T.

2016. Summoning demons: The pursuit of exploitable bugs in machine learning. *CoRR* abs/1701.04739.

[42] Syed, Z.; Padia, A.; Mathews, M. L.; Finin, T.; and Joshi, A. 2015. UCO: A unified cybersecurity ontology. In *AAAI Workshop on Artificial Intelligence for Cyber Security*, 14–21. AAAI Press.

[43] Takahashi, T.; Fujiwara, H.; and Kadobayashi, Y. 2010. Building ontology of cybersecurity operational information. In *6th Workshop on Cyber Security and Information Intelligence Research*, 79. ACM.

[44] Takahashi, T.; Kadobayashi, Y.; and Fujiwara, H. 2010. Ontological approach toward cybersecurity in cloud computing. In *3rd Int. Conf. on Security of information and networks*, 100–109. ACM.

[45] Undercofer, J.; Joshi, A.; and Pinkston, J. 2003. Modeling Computer Attacks: An Ontology for Intrusion Detection. In *Proc. 6th Int. Symposium on Recent Advances in Intrusion Detection*. Springer.

[46] The washington post: Cybersecurity.

[47] Weninger, T.; Zhu, X. A.; and Han, J. 2013. 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*, 579–583. IEEE.

[48] Windrem, R., and William M. Arkin, N. N. Jul 2017. Who planted the fake news at center of qatar crisis?