Gap Finding and Truth Identification from UN Structured Text

Shireen M T
CSE Department
MGM college of Engineering and pharmaceutical sciences
Valanchery, India

Nahan Rahman M K
CSE Department
MGM college of Engineering and pharmaceutical sciences
Valanchery, India

Amina S
CSE Department
MGM college of Engineering and pharmaceutical sciences
Valanchery, India

Abstract: **Textual data mining is an automatic process that uses natural language processing and different AI technologies to extract structured data from unstructured web data. Textual data mining is used in various wide range of domain such as military, cyber security, law enforcement, business. For example, application for text data mining in cyber security have created a scope of threat intelligence that serve the IT business. However, less considered issue in the automatic identification of semantic inconsistencies among different content input. In this paper we are introduce a Gap-finding and truth identification for finding the inconsistency in information and identify the truth about the information. Gap finding is a new inconsistency checking system for recognizing semantic irregularities inside the online protection space, by using feature extraction technique like doc to vector, Latent Semantic Analysis (LSA) and Neural network like Long Short-Term Memory (LSTM). Also, we find Truthiness of information using Term -frequency and Inverse Document Frequency (TF-IDF) and Machine learning classifiers like Support vector machine (SVM)**

Index terms: Cyber security, inconsistency, feature extraction technique, machine leaning algorithm. Support vector Machine, **Term -frequency and Inverse Document.**

I. INTRODUCTION

As the recurrence and refinement of Cyber-attacks keep on rising, so too has developed the cyber threat intelligence (CTI). A several research projects have proposed CTI related systems and approaches [7], [8] to automatically analyze and identify advanced attacks. recommended an approach a way to analyze reported attack scenarios and derive possible defense methods. One commonality among this paper and other proposed CTI method [1]s is a central reliance on publicly available information from blog articles, research papers, some security related reports. Here main question is detection rate and accuracy of web mined information for build CTI system.

The most critical problem while building CTI system is correctness of information and automated inconsistency analysis. This paper explores an approach that automated inconsistency checking and find the which article give the correct information. For example, consider an Information about polymorphic malware that collected different security blog site A and B. One text segment from site A reports that it initially discovered in 1990 September by Mark Washburn, while another text segments from site B it was created on February 1992 by hacker Dark Avenger. While both text segments include polymorphic malware, each mentioned Table 1 different discovered time period. An ideal CTI system detect the inconsistency and correct information i.e., detect which discovery date is most likely to be accurate.

Correctness of information research topic have been proposed in [3],[4],[6],[5] in text mining domains. Some demerits in the existing research paper that cannot apply directly, some reasons are verifying correctness of information already structured, so that cannot apply to evaluating unstructured data, all the data's in the web sources are written in unstructured manner and natural language [13] .Therefore some relationship building steps also needed and some language processing technique like Named entity recognition(NER) and Relation Extraction(RE).Second extraction of security related information from unstructured data need additional formalization that is pre-processing text i.e. removing stop words, tokens, punctuation. Some different terms represent same meaning for e.g., attacking, emails are different words that comes under same technique, *hacking* so we need to refine extracted data, other difficult is to find the semantic relationship for malware name. For example, here Polymorphic malware is a type of malware that constantly changes its identifiable features in order to evade detection also Polymorphic is a cell mean they can exist in more than one form, such as- Lysosomes. Many of the common forms of malware can be polymorphic, including viruses, worms, bots, trojans, or keyloggers so we need to some security terms are related to each other.

Table 1: Example of inconsistent CTI report claims among sources on the first detection date of polymorphic malware.

Source	Part of Articles
A	***From source A reports that it initially discovered in 1990 September by Mark Washburn,***
B	***From source B it was created on February 1992 by hacker Dark Avenger.***
C	***Also, the malware Polymorphic is a cell mean they can exist in more than one form, such as- Lysosomes.***

The last decades there is a rapid growth of social networks [2] also have distribution of latest news, stories, article etc. Anyone can spread any information at any time on many open and always on social media platforms without real- world validation and responsibility, which has resulted cause of spreading of fake news, social spams.

II. PROPOSED SYSTEM

In this paper we propose an orderly way to deal with identifying irregularity in security related data from different articles. Also finding which article give correct information and update that information and discard fake news.

some steps for finding inconsistency: i) data collection, ii) pre-processing of dataset, iii) separate the data based on heading, here there some steps for Separating heading, for example we are using security and non-security-based topics, apply doc2 -vector and LSTM classifier for categorizing the heading then iv) cluster formation, v) Entity Tagging, vi) Relationship building, vii) Malware graph, viii) Information analyzer.

After these steps we can get is any inconsistency present in the article, if present we will find out the real article and update the information in that article. it consists two steps Training and Testing. In Training first, we collect fake and real article from blog, social media, research papers then extract the characters, it consists five steps, i) Web site registration behavior of the publishers ii) Internet site ages of the publisher's iii) Domain ranking iv) Domain Popularity v) probability of news disappearance. Analyze them using TF-IDF and train the dataset using machine language classifier finally we can get fake and real information, real information update and discard fake information.

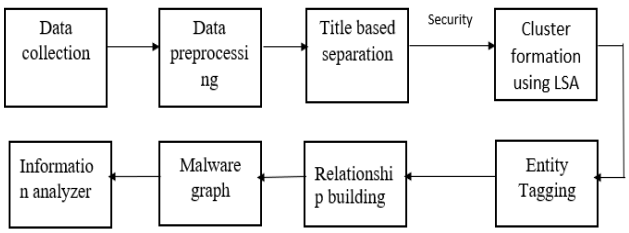

Fig 1: steps for checking inconsistency

Fig 1 illustrate the overall steps for finding inconsistency. In data collection we collect articles from a set of cybersecurity website. For each website explores links.

Table 2 shows a summary of dataset collection., here we collect 1k articles from 50 sources and publication date of the article ranged from 2008-2021.

Article time span	2008-2021
Sources	50
Article	1000
Sentences	105488

After the extraction we categories the title of the article in to two. Showing fig 2

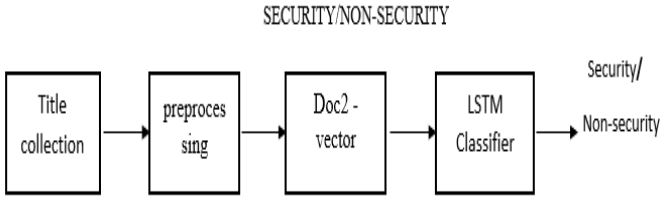

Fig 2: steps for categorizing security/non-security

here we are taking security and non-security-based articles, then second step is pre-processing our title, removing punctuation, stop words, tokenize the text. Doc2-vector model is used to convert word to vector Finally train the model using LSTM classifier in neural network obtaining the security related Heading. In cluster formation we pre-process security related words in data then apply Bag of words method. Latent sematic analysis (LSA) is used to find out semantic relationship. Explained in Table 3

Table 3: Word clustering based on LSA

Cluster	**Frequency**	**Cluster words**
G1	4,436	Polymorphism, malware
G2	1,973	Character, First

After cluster formation here we are doing entity tagging for that we are focus on malware (MW) and date (DA), others (O). The main goal of entity tagging is to extract semantic relationship between the entities. Next step is relationship building here any words related to malware (MW)is grouped one and date (DA) related words also grouped. finally, we get a malware graph and also analyze the information is any inconsistency present or not. Next step finds out the truthiness of the information for that we will train the system and after training we will test the system. For training there some steps are needed. I) first, we collect some fake and real news from online social network, then extract the characters, it consists five steps, i) Web site registration behavior of the publishers ii) Internet site ages of the publisher's iii) Domain ranking iv) Domain Popularity v) probability of news disappearance.

After this we first find out the most important topics in the real and fake news article using Term-Frequency and Inverse Document frequency (TF-IDF) analysis also Subsequently, we investigate the probabilistic LDA subject model to comprehend the distinction or closeness of points between marked fake and genuine news.

In fig 3 explaining some steps to find out the truthiness of information, it consists two steps Training and Testing. In Training first, we collect fake and real article from blog, social media, research papers then extract the characters, it consists five steps, i) Web site registration behavior of the publishers ii) Internet site ages of the publisher's iii) Domain ranking iv) Domain Popularity v) probability of news disappearance.

Analyze them using TF-IDF and train the dataset using machine language classifier finally we can get
fake and real information, real information update and discard fake information.

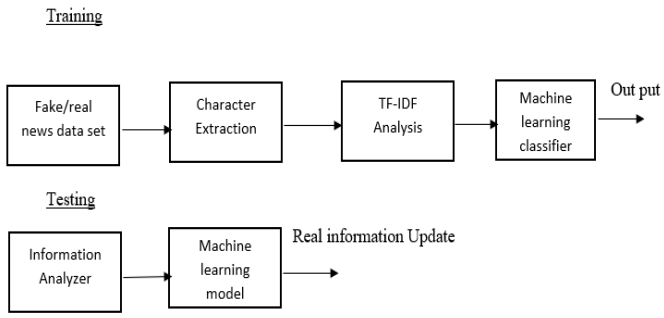

Fig 3: Information Truth Analyzer

III. RESULT ANALYSIS

After the extraction we categories the title of the article in to two. here we are taking security and non-security-based articles, then second step is pre-processing our title, removing punctuation, stop words, tokenize the text.
Doc2-vector model is used to convert word to vector Finally train the model using LSTM classifier in neural network obtaining the security related Heading. Fig 4 shows the training model of LSTM

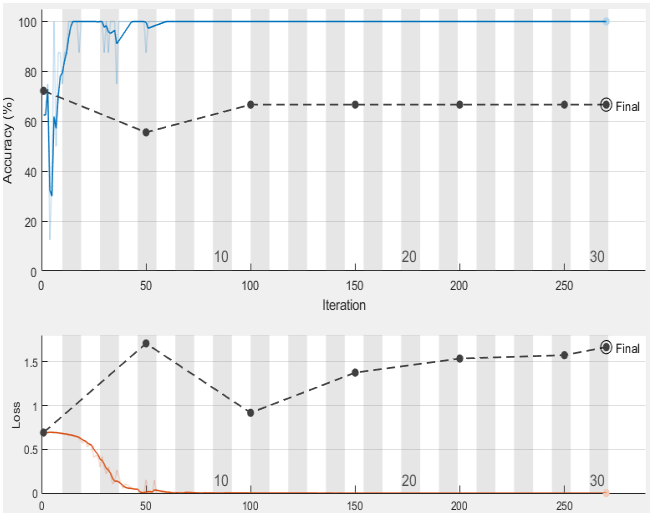

Fig 4: Training Graph of LSTM classifier

After training we will test the inconsistency of data in the article for that we will follow some steps like cluster formation, entity tagging and relationship building finally we will get a malware graph, also analyze any inconsistency in the given information. Fig 5 representing the malware graph, here there some inconsistency in the entity relationship, for example date of malware that is discovered in February and September so here we find out inconsistency in date of data

.

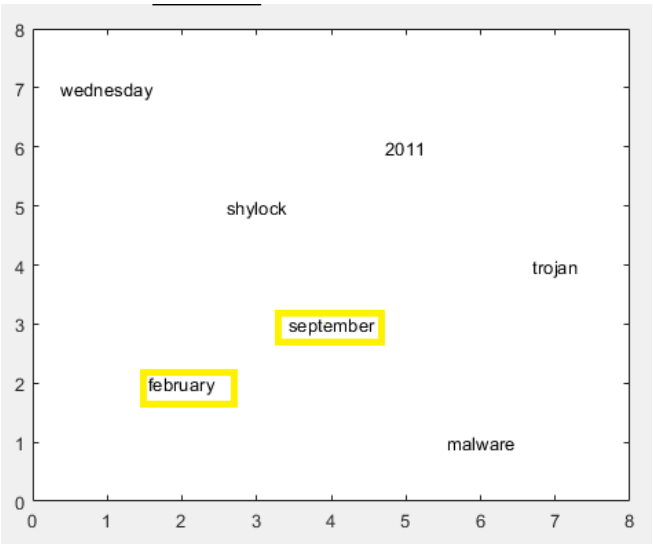

Fig 5: malware Graph

To finding the truthness of information we also measure the domain rankings of Web sites which publish fake or real news.

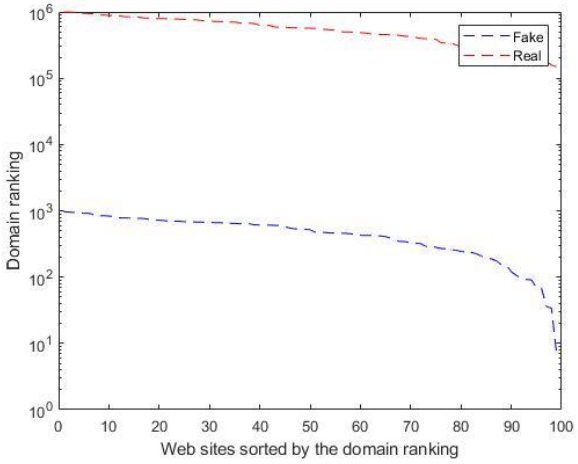

Fig 6: Domain rankings of the fake and real news publishers

Besides of domain Ranking we will use some other characteristics like ages of the domains for the fake and real news, here we characterize the domain age distribution for the fake and real news, that is shown in fig 7 which reveal the very short domain ages for fake news, and the long domain ages for real news. while the fake news driven distributers frequently temporarily register the sites to get out spreading fake news in a brief time of period.

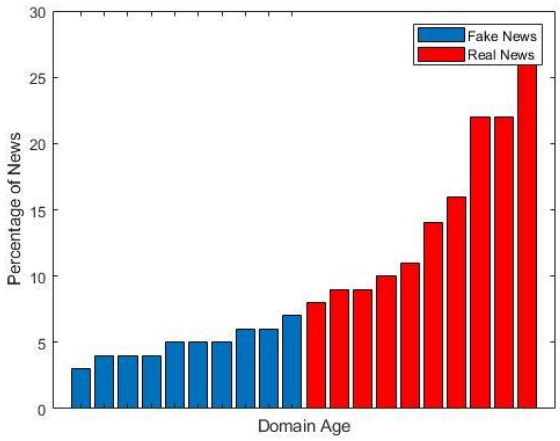

Fig 7: Web site age distribution of the fake news publishers vs. real news publishers

IV. RELATED WORKS

A. *Relation extraction from large pain-text collection*

Text document contain valuable structured data hidden in regular English sentences, by snowball extraction techniques, at each iteration of extraction process snow ball evaluate the quality of these pattern and tuples without human intervention. keep most reliable ones for the next iteration. This technique builds idea of dual iteration pattern expansion. The main disadvantage it is only used in plain text documents.

B. *In web application automatic extraction of IOC*

Indicators of compromise (IOC) found in system log entries or files to identify malware activities.it proposes for the first time an automated technique to extract and validate IOC for web application. The main advantage it can detect attacks and act quickly to prevent breaches from occurring, disadvantage is it create false positive output when we apply a particular environment.

C. *In social networks automated fake news detection*

As of late, the unwavering quality of data on the Internet has arisen as a significant issue of present-day culture. Interpersonal organization locales (SNSs) have reformed the manner by which data is spread by permitting clients to uninhibitedly share content. As an outcome, SNSs are additionally progressively utilized as vectors for the dissemination of deception and lies. The measure of dispersed data and the velocity of its dissemination make it for all intents and purposes difficult to survey unwavering quality in an ideal way, featuring the requirement for programmed trick recognition frameworks.

V. CONCLUSION

For a security related system, we focus on how to extract and analyze open-source information for categorizing and labelling information, and also to ensure the consistency and accuracy of the un structured data and the truthness of information. In this paper, we propose inconsistency checking system, analyzing the structured relations that are extracted from cybersecurity sources. Using a large set of text-mined malware reports from security related sites, here we find syntactic and semantic inconsistencies of information. As fake news and incorrect information continue to grow in online social media, it becomes imperative to gain in-depth understanding on the characteristics of fake and real news articles for better detecting and filtering fake news we build a model using Term-Frequency and Inverse Document Frequency (tf-idf) methods and machine learning classifiers.

REFERENCES

[1]. Text Hyeonseong Jo† Jinwoo Kim† Phillip Porras‡ Vinod Yegneswaran‡ Seungwon Shin† † KAIST ‡SRI International {hsjjo, jinwoo.kim, claude}@kaist.ac.kr {porras, vinod}@csl.sri.com, GapFinder: Finding Inconsistency of Security Information from Unstructured.

[2]. Kuai Xu , Feng Wang, Haiyan Wang, and Bo YangDetecting Fake News Over Online Social Media via Domain Reputations and Content Understanding.

[3]. X. Yin, J. Han, and S. Y. Philip. Truth discovery with multiple conflicting information providers on the web. IEEE Transactions on Knowledge and Data Engineering, 20(6):796–808, 2008.

[4]. B. Zhao and J. Han. A probabilistic model for estimating real-valued truth from conflicting sources. Proc. of QDB, 2012.

[5]. H. Zhang, Y. Li, F. Ma, J. Gao, and L. Su. Texttruth: an unsupervised approach to discover trustworthy information from multi-sourced text data. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 2729–2737, 2018.

[6]. X. L. Dong, E. Gabrilovich, K. Murphy, V. Dang, W. Horn, C. Lugaresi, S. Sun, and W. Zhang. Knowledge-based trust: Estimating the trustworthiness of web sources. Proceedings of the VLDB Endowment, 8(9):938–949, 2015.

[7]. G. Husari, E. Al-Shaer, M. Ahmed, B. Chu, and X. Niu. Ttpdrill: Automatic and accurate extraction of threat actions from unstructured text of cti sources. In Proceedings of the 33rd Annual Computer Security Applications Conference, pages 103–115. ACM, 2017

[8]. X. Liao, K. Yuan, X. Wang, Z. Li, L. Xing, and R. Beyah. Acing the ioc game: Toward automatic discovery and analysis of open-source cyber threat intelligence. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pages 755–766. ACM, 2016

[9]. Z. B. He, Z. P. Cai, and X. M. Wang, Modeling propagation dynamics and developing optimized countermeasures for rumor spreading in online social networks, in Proc. IEEE 35th Int. Conf. on Distributed Computing Systems, Columbus, OH, USA, 2015.

[10]. Z. B. He, Z. P. Cai, J. G. Yu, X. M. Wang, Y. C. Sun, and Y. S. Li, Cost-efficient strategies for restraining rumor spreading in mobile social networks, IEEE Trans. Veh. Technol., vol. 66, no. 3, pp. 2789–2800, 2017.

[11]. C. C. Shao, G. L. Ciampaglia, O. Varol, A. Flammini, and F. Menczer, The spread of fake news by social bots, arXiv preprint arXiv: 1707.07592v1, 2017.

[12]. J. Thorne, M. J. Chen, G. Myrianthous, J. S. Pu, X. X. Wang, and A. Vlachos, Fake news detection using stacked ensemble of classifiers, in Proc. EMNLP Workshop on Natural Language Processing Meets Journalism, Copenhagen, Denmark, 2017.

[13]. G. Salton and M. J. McGill, Introduction to Modern Information Retrieval. New York, NY, USA: McGraw-Hill, Inc, 1986