A DICTIONARY-BASED APPROACH TO IDENTIFYING MALICIOUS MACHINE-GENERATED TEXT

By

Tianyu Wang, PhD

Submitted in partial fulfillment
of the requirements for the degree of
PhD in Computer Science

at

Seidenberg School of Computer Science and Information Systems

Pace University

June 2021

Seidenberg
School of Computer Science and Information Systems
PACE UNIVERSITY
Work toward greatness.

Doctor of Philosophy in Computer Science

Goldstein Academic Center
861 Bedford Road
Pleasantville, NY 10570

Phone: (914) 773-3396
Fax: (914)-773-3533

Dissertation Defense Evaluation Form

Please complete and return form to the Program Coordinator

Student Name Tianyu Wang Student ID# U U00734407

Date: 06/29/2021

Dissertation Title:
A Dictionary-based Approach to Identifying Malicious Machine-generated Text

☑ **Approved.** We hereby certify that this dissertation satisfies the dissertation requirements for the degree of Doctor of Philosophy in Computer Science and has been approved.

☐ **Not Approved.**

	NAME	SIGNATURE	DATE
Committee Advisor/Chair	Chen, Prof. Li-Chiou	Digitally signed by Chen, Prof. Li-Chiou Date: 2021.06.29 11:18:29 -04'00'	
Committee Member	Yegin Genc	Yegin Genc	06/29/2021
Committee Member	Juan Shan	Juan Shan	06/29/2021

Office Use Only

Date Processed_____ By_____

Notes

Abstract

A Dictionary-Based Approach to Identifying Malicious Machine-Generated Text

by
Tianyu Wang

Submitted in partial fulfillment
of the requirements for the degree of
PhD in Computer Science

June 2021

The primary focus of my dissertation is to study how to apply Artificial Intelligence (AI) in cybersecurity to identify malicious machine-generated text. I define the scope of malicious machine-generated text in two forms: 1) A malicious domain name, and 2) misinformation content in social media. I analyze machine-generated textual content and investigate the relationships between the linguistic and information characteristics of dynamic generated content. Thus, my research problem is to distinguish generated textual content between machine (bots) and humans. More specifically, the research problem can be divided into several sub-problems: First, after testing 39 DGA-family domain names and comparing results with other similar research, my method which utilizes N-gram based dictionary features from Alexa and English dictionary outperformed the detection in malicious domain names generated using domain-generated algorithms. Second, I proposed a method which combines a sequence of word frequencies and information entropy of the content to generate features for machine learning algorithms in social media account detection. By detecting such accounts early, this method can stop the spread of false information in a timely manner. Third, I presented an algorithm by incorporating a new similarity-based feature which only extracted from the text content, content-based features, and user-based features in LSTM model. Furthermore, to extend the rational of ensemble learning, I combined two LSTM models by using a conditional meta-classifier in spam detection. The exhaustive experiment showed that this model outperforms all other baselines on an imbalanced dataset and achieves comparable results as a modern model on balanced dataset.

Acknowledgements

This work would not have been possible without the financial support of Seidenberg of Computer Science, Information System Department of Pace University, the National Science Foundation, and National Security Agency. I am especially indebted to Dr. Li-Chiou Chen, Chairman of the Department of Information System, Dr. Yegin Genc and Dr. Juan Shan, who have been supportive of my career goals and who worked actively to provide me with the protected academic time to pursue those goals. I am grateful to all of those with whom I have had the pleasure to work during this and other related projects. Each of the members of my Dissertation Committee has provided me extensive personal and professional guidance and taught me a great deal about both scientific research and life in general. I would especially like to thank Dr. Li-Chiou Chen, the chairman of my committee. As my teacher and mentor, she has taught me more than I could ever give her credit for here. She has shown me, by her example, what a good scientist and person should be. Nobody has been more important to me in the pursuit of this project than the members of my family. I would like to thank my parents; whose love and guidance are with me in whatever I pursue. They are the ultimate role models, who provide unending inspiration.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

With the rise of Artificial Intelligence (AI), automated internet robots, known as bots, have become more universal in the society accomplishing complex tasks in the systems which interact with users and organizations. According to a study in 2017, more than half of all web traffic consists of bots (Glaser, 2017). Some bots are "good", as they are designed to assist in various tasks without any malicious intent. An example of the typical "good" bots is to gather information, such as web crawlers. Another "good" use of bots is automatic interaction with instant messaging or other assorted web interfaces. For example, companies have formed chatbots which can benefit customers, such as a chatbot in airline system allows customers to receive boarding passes, check in reminders, and other information required for a flight. In addition, Google Assistant and Siri allow people to ask questions and get a response using an AI system. These technological advances are positively benefiting people's daily lives.

However, the dynamics of these interactions are impacted by adversarial circumstances, and they are accompanied with detrimental effects in network security (Columbus, 2020). For example, they can be used to launch malicious and harsh attacks against systems or networks. This adversarial effect can be alleviated if people can identify whether the text

is malicious or not. Thus, the detection for malicious text becomes a binary classification problem. However, it is very difficult to model textual data and maintain a robust classification performance. For example, people express their negative sentiments using positive words. This fact allows sarcasm to easily cheat text analysis models unless they are specifically designed to take this possibility into account. Furthermore, word ambiguity is another pitfall that a researcher might face working on a text analysis problem. The problem of word ambiguity is the impossibility to define polarity in advance as the polarity for some words is strongly dependent on the sentence context. Furthermore, due to its unique characteristics, short text classification is deemed to be demanding and challenging. A short text is sparse data which is highly reliant on context so that it becomes problematic to select powerful language features since shared context, and word co-occurrences are insufficient for using valid distance measures. Finally, text data usually requires a special approach to machine learning. This is because text data consist of hundreds of thousands of dimensions (words and phrases) but they tend to be very sparse. For example, the English language has around 100,000 words in common use. However, any given tweet only contains a few dozen of them, which is not quite as sparse. Therefore, whether we can find an adaptive method to analyze textual data with more robust features which fit these circumstances in cybersecurity becomes more challengeable.

In this dissertation, I focus on analyzing adversarial textual contents which are generated by machine. More specifically, I am interested in two types of machine-generated content: 1) URLs with the purpose of promoting scams, attacks, frauds. Such URLs can also be used to avoid blacklist detection in a computer security system. 2) Text that matches the

style of human language reasonably well. Such content can be misused by adversaries, e.g., by automatically generating spams, fake product reviews and fake news, which can seem authentic and fool humans.

1.1 Malicious Machine-generated URLs

Malicious URL, also known as a malicious website, is a common and serious threat to cybersecurity. Malicious URLs host unsolicited content (spam, phishing, etc.) and lure unsuspecting users to become victims of scams (monetary loss, theft of private information, and malware installation), which causes a loss of billions of dollars every year (Sahoo et al., 2017). It is imperative to detect and act on such threats in a timely manner. Traditionally, this detection is done mostly through the usage of blacklists. A blacklist is a basic access control mechanism which allows access to all elements (Emails, URLs, etc.) except those explicitly mentioned. Blacklists of URLs are essentially a database of URLs which have been confirmed to be malicious in the past. However, blacklists cannot be exhaustive, and lack the ability to detect newly-generated, malicious URLs.

Another type of URLs can be used to maintain the communication between infected hosts and the central server(s) in botnet attacks. A botnet, short for “robot network”, is a network of devices infected by malware, and typically, the owner is unaware of the fact that their devices are compromised. The infected devices are under the control of an attacking party, which is called a botmaster, to perform various tasks. A bot could query a pre-defined Command and Control (C&C) domain name or URLs, which resolves the IP address of a server that the malware commands will be received. A new report shows that distributed

denial of service (DDoS) attacks have increased dramatically in the first two quarters of 2018 compared to 2017 (Abrams, 2018). The average size has increased over 500% to 26 Gbps, and the maximum size has increased to 359 Gbps. The increase in attacks is being attributed to large scale botnets being created by attackers using insecure IoT devices (Nexusguard, 2018). To that end, many botnet detection systems use blacklists of C&C domains or URLs to halt their traffic. In turn, botmasters have begun generating domain names dynamically to pass through blacklists. A technique called "Fast flux" is used by botnets to hide phishing and malware delivery sites behind an ever-changing network of compromised hosts acting as proxies. Specifically, they generate several domain names automatically using domain generation algorithms (DGAs), and they only register a subset of these auto generated domain names for actual use (Plohmann et al., 2016). Domain generation algorithms (DGAs) can generate various random domains dynamically, and hence, the static domain blacklists become ineffective to prevent malicious attacks by botnets. Various families of recent botnets, such as Necurs, Dridex, Kraken and Mirai, have used DGA to establish communication with bots and malicious software. Therefore, the detection of DGAs domain names becomes crucial due to its dynamic characteristics.

1.2 Malicious Machine-generated Misinformation Content

Social media provides significant improvements on how information is diffused into systems and businesses. Simos (2015) reported that by 2020, it was estimated that 1.7MB of data will be created every second for every person on earth. In this circumstance, a large portion of data is generated by machine or scripts, which is also called "bot". However, the

quality of the information shared and the intent for information sharing play a dominant role on whether their impact is positive or not. For example, recent studies reviewed the use of social media for bullying among teenagers, also known as cyberbullying (Kowalski & Limber, 2007). Similarly, social media can be exploited for the widespread “fake” or low-quality information (Ciampaglia & Menczer, 2018). Finding episodes of ill-intended or incorrect information is a complex task, and oftentimes, it requires the understanding of the content being shared.

The spread of false or low-quality information is typically handled with automated social media accounts, controlled by bots in platforms, such as Twitter. Such accounts have recently raised significant attention due to the potential consequence in political realm. NBC News has published a database of more than 200,000 tweets which Twitter has tied to "malicious activity" from Russia-linked accounts during the 2016 U.S. presidential election (Popken, 2018). In addition, it was predicted that Russia might influence the 2018 US midterm elections (Stukal et al., 2017). Therefore, identifying whether a social media account is authentic (manually managed) or operated by spam bots as early as possible is crucial for public information security. Twitter, as one of most popular social media platforms, has become the main target for spamming activities, and the accounts and spams are usually generated by machines or programs. Most of the current spam detecting methods in Twitter focus on account level due to limited information in a tweet (Madisetty & Desarkar, 2018). However, spammers can create a new account and begin posting new spam tweets. Therefore, spam detection at the message level is very challenging and highly in demand.

1.3 Research Problems

After assessing the characteristics required by an efficient method in terms of speed, coverage, reliability, and ease of use, I exploit the fact that URLs, domain names, and social media information belong to the same category of textual information. Thus, I define the scope of malicious machine-generated text as two forms: 1) A malicious URL or domain name, and 2) A piece of malicious text misinformation content in social media.

I analyze machine-generated textual content and investigate the relationships between the linguistic and information characteristics of dynamic generated content. All these problems lend themselves to classification problems, such as DGA vs. non-DGA, bot vs. non-bot, and spam vs. human. Thus, I chose to comb the complexity of AI techniques with Nature Language Processing technique and used information theories to solve them. Hence, my research problems can be narrowed down into several dimensions.

Research Problem 1: Can we distinguish whether a URL or domain name is generated by bot or human?

More specifically, the goal of this research is to detect DGA domain names or URLs. This is a string-level text binary classicization problem. Each input entry would be classified as either a positive case or a negative case, which are defined as the following:

- **Positive**, or "**DGA**": the domain is generated by a domain name generation algorithm; and
- **Negative**, or "**legit**": the domain is legitimate and, therefore, it is created by humans, such as system administrators.

Follow the text-string level classification task, another question arose that whether I can extend this problem into a more universal circumstance. I can release the constraint in short URLs format to a more general context in social media.

Research Problem 2: Can we distinguish whether a user or an account in social media is a bot or human?

The purpose of this research is to detect whether social media accounts are utilized by machine (spambot) by analyzing only the aggregated content of the contents without associated meta-data. The classification cases are defined as follows:

- **Positive**, or "**spambot**": the social media account is managed by a bot or machine.
- **Negative**, or "**human**": the social media account is managed by a human.

Additionally, after considering the account-level spam detection, I plan to further extend a more general condition to the message-level. However, detecting spams or bots at message-level or tweet-level becomes particularly difficult, because they are relatively rare in real world and are usually identified in very large and complex datasets, which are generally referred to as a class imbalance problem. An imbalanced classification problem is an example of a classification problem where the distribution of examples across the known classes is biased or skewed. Imbalanced classifications pose a challenge for predictive modeling, as most of the machine learning algorithms used for classification were designed in the assumption of an equal number of examples for each class.

Research Problem 3:

(1) Can we introduce a new measurement which is only extracted from text, and improve performance in spam detection for imbalance data?

(2) Does applying this method also lead to further improvement of the classification results for the balanced data?

The goal of this research is to introduce a new measurement in feature extraction on textual data and compare its effectiveness and resilience in text classification on both balance and imbalance data. The classification classes are defined as follows:

- **Positive**, or "**spam**": the post of social media (such as a tweet) is a spam.
- **Negative**, or "**non-spam**": the post of social media is not a spam.

1.4 Dissertation Organization

This dissertation is organized as follows. The next section explains literature. The proposed detection methodology is presented in section 3. Three completed researches are shown in section 4, section 5, and section 6, respectively. At last, I summarize major contributions in section 7.

Chapter 2 Literature Review

I have reviewed and categorized qualitatively current detection of malicious machine-generated text which has appeared in the literature. The characterization is based on the generation mechanism and detection approaches as the performance tradeoffs and deployment costs of a detection are dependent on them. A generation mechanism refers to the approach that schemes and tools use to generate malicious text. A detection approach refers to the method which identifies malicious text.

2.1 Generation Approaches for Malicious Machine-generated Text

I briefly discuss some of the adversarial text generation techniques. Here I discuss a few more proposals with particular emphasis on the diversity of the intention. I organize the generation of malicious text based on the domain in the following.

2.1.1 Command & Control Server's URLs

One of branches of machine-generated text is used to generate domain names for the access to command and control server. This category of text is used as a link to redirect to a website domain names or web servers instead of meaningful content. Most of these attacking techniques are realized through spreading compromised URLs, and the spreading of such URLs forms a critical part of the attacking operation. Domain generation algorithms (DGAs) can produce many random domains dynamically, which causes the static domain blacklists to become ineffective in preventing malicious attacks by botnets. Those malicious URLs usually arise new sophisticated techniques to attack and scam users.

Such attacks include rogue websites which sell counterfeit goods, financial fraud by tricking users into revealing sensitive information which eventually lead to theft of money and identity, installing malware in the user's system or performing botnet attacks.

Nowadays, many malicious URLs are using random generators or domain name generation algorithms (DGAs) to create multiple domain names and register only a subset of these domain names for actual use, which is a process commonly referred to as "domain fluxing". With domain fluxing, attackers query each of the auto-generated domain names until one of them resolves to an IP address. The resolved IP-address obtained is then used to host the command-and-control (C&C) activities (Yadav et al., 2010) (Yadav et al., 2012). In short, the use of DGA allows attackers and victims machines to contact the same domain at a given time, so the exchange of information is possible at least temporarily. These connections points are hard to predict for parties who do not have access to the algorithm. Recent examples of malware attacks with DGA features are variations of the Mirai botnet (Rodriguez, 2016). Mirai botnet has become one of the most widespread malwares which infected various Internet of Things (IoT) devices. In the meantime, Mirai botnet can turn networked devices running Linux into remotely controlled "bots" for large-scale network attacks. It primarily targets online consumer devices, such as IP cameras and home routers, and it has been used in some of the most significant and most disruptive attacks (Krebs, 2016; Mendez Mena et al., 2018).

Apart from the characteristics of seeding, four different generation schemes emerged during the analysis:

1) Arithmetic-based DGAs calculate a sequence of values which either have a direct ASCII representation usable for a domain name, or designate an offset in one or more hardcoded arrays, constituting the alphabet of the DGA (Yadav et al., 2010). They are the most common type of DGA.

2) Hash-based DGAs use the hexdigest representation of a hash to produce an Algorithmically Generated Domain (AGD). They are identified DGAs using MD5 and SHA256 to generate domains (Yadav et al., 2010).

3) Dictionary-based DGAs will concatenate a sequence of words from one or more wordlists, resulting in a less randomly appealing and thus more camouflaging domains (Curtin et al., 2019; Geffner, 2013; Plohmann et al., 2016). These wordlists are either directly embedded in the malware binary or obtained from a publicly accessible source.

4) Permutation-based DGAs derive all possible AGDs through permutation of an initial domain name (Yadav et al., 2010).

2.1.2 Spam Contents

Subsequently, one big source of adversarial text which was generated by machine is spam. Spam in a computer science field refers to the unwanted or unsolicited messages sent or received electronically by means of e-mail, instant messenger, social networks, etc. As evident from the definition of the term, spam is intended for malice and usually accounts for a viable but fraudulent source of income for some individuals or organizations. An individual involved in sending such spam messages is generally termed as a “spammer”.

Spam is inevitable in almost all forms of online communication today and is known to hamper the productivity of the medium on which it appears.

Due to differences in the characteristic features of social websites from the features of usual e-mails and social networks, spam fighting is multifaceted and challenging. It is important to know the existing spamming tools and techniques. A good attempt in this direction can be found in research conducted by (Stern, 2008). The article has focused on e-mail spamming and described the operation of three specific technologies: Dark Mailer, Send Safe and Reactor Mailer. (Cournane & Hunt, 2004) discussed mass (or bulk) e-mail software which allows a spam source to create a template message to be sent to a list of recipients. A research has explored the spamming tools and techniques (Chaitanya et al., 2012). It identifies various spamming techniques used by URL shorteners, which makes spamming more prevalent in emails, social networks, and reviews.

More concerns have been raised regarding the possibility of manipulating public opinion through social media (Enli, 2017). Social media have been proved effective in influencing individuals, their beliefs and behaviors (Aral & Walker, 2011; Centola, 2011; Mønsted et al., 2017), which is a particularly problematic fact. These concerns have been later proved well-grounded by several scientific studies, which highlighted a variety of manipulation strategies and related contexts where such forms of abuse can take place. Thus, one way to manipulate social media is by using social bots, algorithmically controlled accounts that emulate the activity of human users but operate at a much higher pace (e.g., automatically producing content or engaging in social interactions), while successfully keeping their

robotic identity undisclosed (Ferrara et al., 2014; Hwang et al., 2012; Messias et al., 2013; Varol et al., 2017a).

Lastly, with the development of artificial intelligence, more machine-generated text is produced by text generative models (TGM) which mimics the style of human language, especially in terms of grammaticality, fluency, coherency, and usage of real world knowledge (Brown et al., 2020; Deng et al., 2020; Radford et al., 2019; Zellers et al., 2019), However, TGMs can have unfortunate uses by adversaries for malicious purposes, such as generation and spread of fake news (Brown et al., 2020; Zellers et al., 2019), generation of fake product reviews (Adelani et al., 2020), and spamming/phishing (Seymour & Tully, 2018).

2.1.3 Categorization of Malicious Text Generation

Depending on the intention, scope on which it is posted, and generation schemes, malicious machine-generated text can be broadly classified into various categories specifically. Table 1 shows a brief comparison of categorization of literature review mentioned above.

Table 1. Categorization of Malicious Text Generation

Intention	Scope	Literature	Generation Schemes & Tools
Command & Control	Bots	(Krebs, 2016; Mendez Mena et al., 2018; Rodriguez, 2016; Yadav et al., 2010; Yadav et al., 2012)	Arithmetic-based DGAs
		(Yadav et al., 2010)	Hash-based DGAs
		(Curtin et al., 2019; Geffner, 2013; Plohmann et al., 2016)	Dictionary-based DGAs
		(Yadav et al., 2010)	Permutation-based DGAs
Spam	Email	(Cournane & Hunt, 2004)	HailStorm, Cyclone, Popup Sender
		(Stern, 2008)	Dark Mailer, Send Safe, Reactor Mailer
	Social media	(Chaitanya et al., 2012)	URL shorteners
		(Ferrara et al., 2014; Messias et al., 2013; Varol et al., 2017a)	Spambot
		(Hwang et al., 2012)	Socialbot Network
		(Seymour & Tully, 2018)	LSTM
	News	(Zellers et al., 2019)	Grover
		(Brown et al., 2020)	GPT-3
	Reviews	(Adelani et al., 2020)	GPT-2 NLM