Dark Web Hidden Service Content Classification using Quantum Encoding

Ashwini Dalvi [1*], Soham Bhoir [2], Faruk Kazi[1], S G Bhirud[1]

Veermata Jijabai Technological Institute, Mumbai, India
[1]* *aadalvi_p19@ce.vjti.ac.in*
[1] *fskazi@el.vjti.ac.in*
[1]*sgbhirud@ce.vjti.ac.in*

K. J. Somaiya College of Engineering, Mumbai, India

[2] soham.bhoir@somaiya.edu

ABSTRACT

Researchers and security professionals consider data collected from the dark web as one of the measures for proactive cyber security. Therefore, classifying dark web content with approaches ranging from machine learning to deep learning is researched extensively in the literature. Still, particular challenges remain with classifying dark web hidden services, for example, the limitation of the dataset to label hidden services and the requirement of substantial computing and storage resources to manage raw and unlabelled dark web data.

The proposed work presented a quantum encoding based approach to categorizing Tor hidden services. First, the dark web crawler crawled the Tor dark web to fetch hidden services. The classical model classifies hidden services into twelve categories. The twelve categories include law and government, forum, streaming services, social networking sites, food, travel, games, health and fitness, education, computer and technology, E-commerce, and business/corporate. The keyword dataset to label hidden services is created using scraping and cleaning surface webpages of each twelve categories. Thus authors address the first challenges of labelling hidden service data. The dark web crawler can crawl hundreds of hidden services in a stipulated time. The crawled data comprised HTML pages and associated files. Analysis of such vast data is demanding on computational resources.

Recently, quantum computing is coming up as an alternative to classical models. Therefore, the authors proposed a quantum model to categorize crawled Tor hidden services with lesser time and minimize memory consumption. The proposed quantum model used Universal Sentence Encoder to encode classical data into quantum data as probabilities in the proposed work. The quantum circuit receives these probability values. Further, the quantum model applies the softmax function before comparing the output to actual category labels.

The result shows the output of categorizing hidden services using a customized category-defining dataset with quantum encoding. Finally, the chapter compares the time and memory consumption between the classical and quantum model.

Keywords: Dark web, Tor, Hidden service, Onion Services, Quantum Encoding, Quantum Circuit, Quantum Computing for cyber security

1. INTRODUCTION

In order to combat security vulnerabilities, security professionals develop a proactive security posture by identifying specific hacker methods and techniques. In the last five years, security attacks like Wannacry and Log4j provoked security professionals and researchers to recognize that implementing a security incidents inventory is the proactive way to mitigate cyberattacks. The objective of proactive cyber security is to identify and correct weak security layers as early as possible, as well as develop strategies for detecting threats in advance.

Using data from public sources, Craig A. N et al. compared 22 cybersecurity companies and 27 cybersecurity solutions to comprehend the most proactive cyber security practices [1]. The top proactive practices followed were auditing, data mining, and analysis. Meland P. H et al. discussed various cyber security indicator points. These points are categorized based on the data source, type, and category [2]. For example, researchers considered dark web marketplaces as emerging remote data points for proactive cyber security measures. Dark Web Monitoring enables organizations to be vigilant of cybercriminals with proactive intelligence. Hackers often conduct several malware and ransomware campaigns on the dark web and the sale of stolen personal and business IDs. Extremists' communication or terrorists' activities are significant concerns related to the dark web, but potential cybersecurity threats are also likely to emerge from the dark web.

Dark web access is facilitated through a specialized network, typically the Tor network. Tor provides anonymity without disclosing the IP addresses of the host and client by going through a network of multiple servers and encrypted networks. The service referred to as onion services (OS), also known as Tor Hidden Service (HS), facilitates anonymity. The following text will use the terms onion service and hidden service interchangeably. A client cannot learn the IP address of the server providing the onion or hidden services. The anonymity of the dark web makes it impossible for law enforcement agencies to identify online transactions on the dark web. The researchers Arnold N. et al. presented a tool to use dark web data as a source for cyber threat intelligence [3]. Shakarian, P. discussed that the dark web could effectively impact cyber threat intelligence by automating the intelligent resources to obtain and analyze dark web information to predict cyberattacks [4].

Researchers pursue cyber threat intelligence (CTI) framework to perceive proactive cyber security. Thus, Zhang et al. developed DWTIA, a Dark Web Threat Intelligence Analysis Platform, to facilitate the dark web analysis for crime and criminal information [5]. Since IP addresses are used for tracing networks, the DWTIA framework did not identify cyber criminals because of the dark web's anonymity. Besides collecting data from more than 8,000 dark web sites, it also provides or uses the OnionScan dark web crawler.

To provide cyber threat intelligence (CTI), Samtani et al. used Diachronic Graph Embedding Framework (D-GEF) technologies to identify trends and tool functionality in online hacker forums [6]. A Graph-of-Words representation of threats emanating from hacker forums was the basis for D-GEF. Jeziorowski et al. analyzed dark marketplace images [7]. The image-based intelligence was collected by identifying reused images, and further using image metadata hashes and image hashing techniques, the computational overhead associated with the process was minimized. The research facilitated by identifying dark marketplaces from well-established dark net market archives resulted in the dark web marketplace vendor profiling. Such dark web investigations can deanonymize anonymous paraphernalia sellers by studying the aliases associated with vendors who conduct business in multiple marketplaces. The top vendors, top markets, and top hash analysis results were identified based on the investigation results. In this study, machine learning-based classification techniques or methods validated the outcomes of multimodal DVP.

According to Meland et al., the dark web is experiencing a rise in ransomware services (RaaS) [8]. Cybercriminals or malicious users on the dark web demand ransom payments to release infected digital assets. The main objective of this study was to examine RaaS and the associated value chains associated with the dark marketplace. According to Meland et al., if perpetrated by an experienced attacker, ransomware bought from the dark web is a potent threat.

In a recent study, Koloveas P et al. discussed the need to develop investigation frameworks for the dark web's Internet of Things (IoT) threat vectors [9]. The work proposed three components: a focused crawler, a forum-based crawler, and a Tor crawler. First, a focused web crawler was designed to identify new resources and gateways for CTI. Next, positive and harmful web pages were classified using a Support Vector Machine (SVM) classifier. In order to locate the initial seed links, the crawler used a user-provided query. Finally, in-depth crawlers focus on forums whose topics of discussion are relevant by examining all the links in the forum. Regex-based link filters, however, monitored forums selectively within and across them.

Various research approaches were studied to analyze content on the dark web. These approaches include analyzing drug market data and hacker forums, among others. Content hosted or offered in hidden services is most frequently the subject of research. However, dark web hidden services (HS) are difficult to identify without opening them. Also, services on the dark web are not indexed, which makes it challenging to determine what content onion service or hidden service contains. Therefore, researchers proposed ways to identify the type of hidden service using text analysis. Although the dark web hidden services consist of unstructured text data after stripping HTML tags from hidden service web pages, text analysis has become increasingly important for seeking useful insights from hidden services. For example, text mining can classify hidden services into different categories. With supervised machine learning, web pages require a labelled dataset to categorize them into different categories. However, the onion service dataset qualifies as an unlabelled dataset. Therefore, the researchers developed custom labelled datasets by identifying the content of hidden services in collaboration with experts. For example, Al Nabki M W et al. released dark web data set comprised of 26 classes to classify hidden services hosting illegal activities [10]. A relatively early attempt by Guitton C. in 2013 also manually categorized 1171 hidden services into 23 categories [11]. The purpose of classifying dark web content is to use further rank the content to offer the better result to law and enforcement agencies. Based on content analysis, Faizan, M et al. proposed a methodology to rank drug-related hidden services [12]. Thus, the researchers attempted different ways to classify dark web data for different purposes.

Researchers collect the dark web data either with a customized dark web crawler or an open-source web crawler. Crawling the dark web results in vast amounts of raw textual data presented in hidden services, but classifying them is time-consuming and costly due to the need for human judgment. In recent times, quantum computing has gained attention because it can achieve exponential speedups over conventional machine learning. The proposed chapter implements a quantum encoding approach to categorize hidden services into predefined categories.

The chapter offers the following novel contributions:
Creation of a customized dataset of different categories to label hidden services
Implementation of quantum encoding to classify Tor hidden services

ORGANIZATION OF CHAPTER

The organization of the chapter includes related work in section II. Section III covers methodologies and result discussion, and the chapter concludes with limitations and future scope.

2. RELATED WORK

Search Engines do not index the Tor hidden services, so a manual effort is required to access hidden services. Researchers have been studying Tor dark web extensively; however, Huete Trujillo et al. cited research for Tor hidden services [13]. The study mentioned six significant areas of hidden services – classifying the content of hidden services, analyzing the security and performance of hidden services, discovering and measuring approaches for hidden services, and changing the design of hidden services. In this work, the authors will focus on classifying Tor hidden services.

Thus it is comprehended from research that crawled hidden services will be unlabeled data and dataset to labelled content of hidden services created by manual or automatic labelling. The proposed study created an automated dataset for Tor hidden service labelling.

Collecting Hidden services for categorization

The dark web investigation mechanism involves a crawling module, link collection and content labelling modules. Using an adaptive learning algorithm, Zhao, F et al. proposed an intelligent crawler capable of selecting features online and automatically constructing link rankers [14]. The crawler consists of categorization websites to exclude unrelated websites. The researchers mentioned the crawler harvest at a higher rate than other variants of deep and dark web crawlers. The dark web crawlers collect data for specific use cases like drug markets and child sexual exploitation. Frank R et al. developed a dark web crawler to analyze child exploitation networks extractor (CENE) on the dark web [15]. Zulkarnine A T et al. modified the capability of the CENE Tor crawler to collect data from the surface web as well [16]. The present work collected data with a crawler capable of crawling surface and dark web [17].

Labelling dark web data for categorization

The onion services are not available publicly; thus, collecting large amounts of onion services to train data for the supervised learning model is difficult. Therefore, the dark web hidden service classifier typically learns from a training dataset consisting of observations related to a particular labelled dataset derived from empirical data or acquired from experts.

The research has focused on manually labelling crawled dark websites and using them as training corpora for automated crawls of the Tor dark web. Dong F et al. manually labelled eight thousand samples to identify thirty-five new cyber threats [18]. Dalins J et al. crawled webpages from different Tor domains and trained the machine learning model by manually labelling 4.000 Tor pages [19]. He S et al. generated a dark web data set comprising 4,851 .onion sites categorized manually [20]. The authors used publicly available legal documents for the classification model training. Such an attempt introduced another possibility for defining illegal activity categories which were not labelled earlier. Mahor V et al. extended the approach of curating a dataset for categorizing cyber threats concerning cyber-physical systems [21].

To determine the characteristics of each category, Takaaki, S., & Atsuo, I. manually analyzed 300 websites and extracted characteristic keywords from each website [22]. As part of this process, keywords deemed most likely to fit within particular categories were enumerated and considered simultaneously. As a result, each category was defined based on ten or more keywords. As a result, the authors identified six categories based on the top page text downloaded from the onion domain.

Researchers also attempted to label dark web data with machine learning. Kobayashi H et al. collected data from the dark web and proposed a system to perform morphological analysis using natural language processing [23]. Authors built a knowledge base mechanism to automatically identify five crime categories and determine threat levels based on data collected from the dark web. Kinder A et al. scraped HTML from the onion service with a predetermined list of keywords to identify illicit sites and categorize them based on their crime [24]. Ghosh S et al. developed 'ATOL –Automated Tool for Onion Labelling' [25]. The ATOL generated new keywords for different categories of onion services with expert-provided keywords.

Buldin, I. D., & Ivanov, N. S. presented work on dark web text classification in the Russian language [26]. Moraliyage, H. at al. extends dark web text labelling with multimodal deep learning to classify multiple onion service categories [27].

The work discussed concerning dark web analysis requires an efficient mechanism to collect and categorize onion services. The present work investigates the advantage of quantum encoding g to categorize hidden services.

Introduction to Quantum Computing

Information and communication technologies will benefit from quantum computers, which are highly powerful and secure. Quantum computing is evolving technology that uses quantum mechanics to solve problems difficult for classical computing. Quantum computers can process exponentially more data than classical computers because they use probability instead of just 1s and 0s. In quantum computing, qubits, a basic memory unit, are created using electrons' spin or photon orientation of physical systems. Quantum superposition refers to the fact that physical systems can exist simultaneously in many different configurations. The quantum entanglement phenomenon is also capable of linking qubits inseparably. Consequently, each qubit can represent a unique thing at the same time.

For example, a classical computer uses eight bits to represent numbers between 0to 255. Quantum computers can, however, represent all numbers between 0 and 255 simultaneously using eight qubits.

Fault-tolerant quantum computers can resolve problems like integer factorization and unstructured database searches by utilizing millions of qubits with low error rates and long coherence times. Even though the experimental progress toward realizing noisy intermediate-scale quantum (NISQ) computers could take decades, noisy intermediate-scale quantum computers are currently available. Noise qubit computers use hundreds of uncorrected quantum bits, resulting in imperfect calculations within a limited period of coherence. Researchers have proposed multiple algorithms to achieve quantum advantage with these devices.

Quantum computing demonstrates an obvious advantage over classical computing. Quantum computers have the potential to surpass the supercomputers of today. However, researchers still investigate which type of problems from classical computing to solve through quantum computing. Further, researchers examined quantum machine learning.

Quantum computer processes data on a quantum level. Quantum software progressed to develop machine learning faster than conventional computers. The first step in quantum machine learning is to load conventional data into the states of the qubits. Quantum states are attained by encoding or embedding quantum data. Quantum Machine Learning algorithms (QML) rely heavily on classical data encoding to perform effectively and efficiently. Quantum machine learning involves three phases – encoding, process and measurement.

Three phases of quantum machine learning are encoding, processing and measurement—the following points offer a brief mention about three points.

Encoding: The process of loading classical data into a quantum state.

Processing: The embedded input, which will be a variational circuit or a quantum routine, is processed by the quantum device at this point.

Measurement: This stage measures the predicted result, subsequently forming the forecast for QML.

Quantum computing for proactive cybersecurity

The quantum computation concept has also influenced many scientific studies in computer science, notably computational modelling, cryptography theory, and information theory. Quantum computers can have either a positive or negative impact on the security of information. Many researchers have evaluated quantum computing's benefits in cybersecurity. Njorbuenwu, M et al. discussed several fields that might benefit from using a quantum computer [28]. Laxminarayana N et al. presented a study on the combination of principles of quantum mechanics and neural networks to train intrusion detection systems for healthcare systems [29]. Researchers demonstrated the proposed algorithm on the KDD99 dataset.

Researchers explored the role of quantum computing in mitigating domain-specific security. To emphasize that quantum computing will be a potential solution to strengthen cybersecurity, Ko K. K. & Jung, E. S. described quantum computing-based implementations of existing AES and modified AES algorithms [30]. Tosh D et al. proposed using quantum cryptography to encrypt communication between sensors and computers to secure cyber-physical systems [31].
Ali, A explored the possibility of combining quantum computing with classical computing [32]. Suryotrisongko, H., & Musashi, Y. proposed the novel hybrid quantum-classical deep learning model for domain generation algorithms (DGA)-based botnet detection [33].
Researchers are also investigating whether quantum mechanical principles can be incorporated into machine learning problems to improve the solution. Abohashima Z et al. summarized the most recent research findings in quantum machine learning [34]. The authors proposed to propose a quantum classification scheme as well as a quantum encoding scheme.

3. PROPOSED APPROACH

The proposed study attempts to classify crawled hidden service websites, legal and illegal, into various categories. One publicly available dark web dataset is the DUTA dataset, which consists of over 10,367 onions services manually labelled in twenty-five categories. The work of the DUTA dataset motivated authors to categorize onion services in general categories. The authors implemented a content-based classification approach to create the dataset. Further, the authors attempted quantum encoding to classify hidden services into different categories.

Figure 1 depicts the proposed methodology for categorizing onion services with quantum encoding.

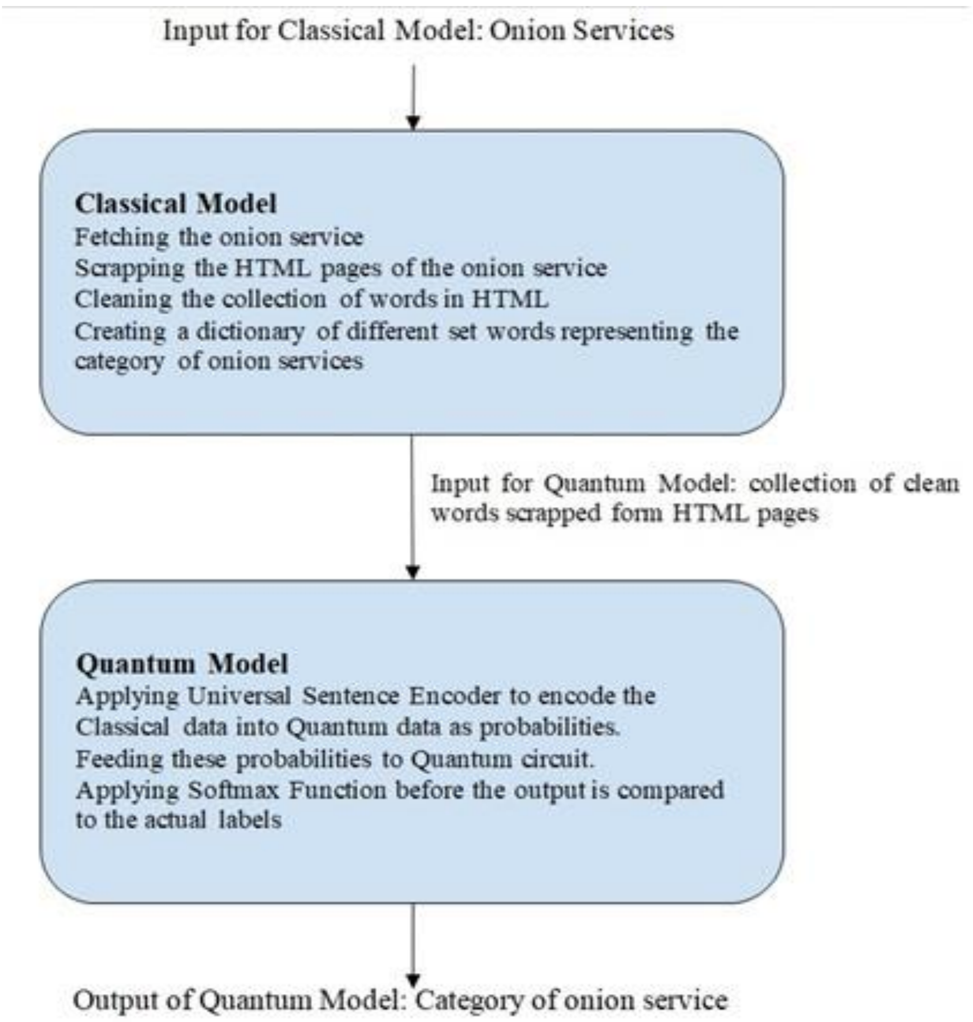

Input for Classical Model: Onion Services
Classical Model
Fetching the onion service
Scrapping the HTML pages of the onion service
Cleaning the collection of words in HTML
Creating a dictionary of different set words representing the category of onion services
Input for Quantum Model: collection of clean words scrapped form HTML pages
Quantum Model
Applying Universal Sentence Encoder to encode the Classical data into Quantum data as probabilities.
Feeding these probabilities to Quantum circuit.
Applying Softmax Function before the output is compared to the actual labels
Output of Quantum Model: Category of onion service

Fig 1 Proposed Methodology

Classical model to categorize onion services

In order to predict each class correctly, it is essential to have a balanced dataset. Therefore, the authors scraped over a hundred surface websites to create a balanced dataset of keywords in twelve categories. The twelve categories include law and government, forum, streaming services, social networking sites, food, travel, games, health and fitness, education, computer and technology, E-commerce, and business/corporate. Data preprocessing involved the removal of HTML tags and non-ASCII characters from scraped HTML pages. Text preprocessing functions – stemming and lemmatization performed the cleaning of raw text data.

Further, the text was categorized into different domains using the TF-IDF approach. Then, the authors trained supervised learning models like Random Forest, LinearSVC, Multinomial Naïve Bayes and Gaussian Naïve Bayes, using a set of keywords of different categories. Finally, the k-fold cross-validation score was measured to evaluate the training models' accuracy. During the machine learning process, cross-validation is an effective measure for identifying overfitting that determines the model's performance with unknown data. The table 1 shows the five-fold cross-validation accuracies of supervised learning models.

Table 1 Five-fold cross-validation scores for supervised learning models

Model Name	**Fold Index**	**Accuracy**
Random Forest	0	0.719858
Random Forest	1	0.751773
Random Forest	2	0.716312
Random Forest	3	0.736655
Random Forest	4	0.679715
LinearSVC	0	0.858156
LinearSVC	1	0.932624
LinearSVC	2	0.939716
LinearSVC	3	0.903915
LinearSVC	4	0.879004
Multinomial Naïve Bayes	0	0.812057
Multinomial Naïve Bayes	1	0.879433
Multinomial Naïve Bayes	2	0.872340
Multinomial Naïve Bayes	3	0.882562
Multinomial Naïve Bayes	4	0.818505
Gaussian Naïve Bayes	0	0.702128
Gaussian Naïve Bayes	1	0.762411
Gaussian Naïve Bayes	2	0.780142
Gaussian Naïve Bayes	3	0.754448
Gaussian Naïve Bayes	4	0.644128

The accuracy of Linear SVC was 0.94. Thus, LineraSVC was selected as a classical supervised learning model.

Quantum Model to categorize onion services

TensorFlowHub has released a pre-trained model for Google's Universal Sentence Encoder (USE). As mentioned in [35], it converts natural language texts into high-dimensional vectors to perform text analysis. USE is designed specifically for material that exceeds the length of a word, such as paragraphs, sentences, or phrases. Various data sources and workloads train the USE model to accommodate a wide range of natural language comprehension tasks. Input English text of variable length produces a 512-dimensional vector.

According to the transfer learning paradigm, the USE model could be a component of a more extensive network. The USE model is used just as-is with its parameters fixed or the model fine-tuned the parameters to optimize them. The output of the USE model is then transferred to subsequent levels to train the network as a whole for a specific downstream job. Document categorization is most likely the most intuitive.

Authors transformed the abstract of a collection of words retrieved from hidden services belonging to distinct categories to sentence embeddings and attached a dense layer with two outputs. Each output node corresponds to a single category. Quantum computer calculations are noisy. Therefore, it is necessary to mitigate errors after a quantum computer calculation.

A softmax function is applied before the output. The softmax function converts a vector of K real values into the vector of K real values equalling 1. The softmax transforms input values, positive, negative, zero, or greater than one, into values between 0 and 1 as probabilities. Although it will always lie between 0 and 1, the softmax translates small or negative inputs into small probabilities and large or positive inputs into high probabilities.

The mathematical definition of the softmax function is depicted in figure 2.

$$\sigma(\overline{Z})i = \frac{e^{Zi}}{\sum_{j=1}^{K} e^{Zi}}$$

Fig 2 Softmax Function

Table 2 discusses the softmax function in detail.

Table 2 Details of Softmax Function

$\overline{Z}$	The softmax function's input vector is built from (Z0, ... Zk)
Zi	The softmax function's input vector contains all of the Zi values, and they can all have a real value of either a positive, zero or negative sign. The softmax is required because, for instance, a neural network's output vector may be (-032, 4.12, 6.47), which is not a legitimate probability distribution.
e^{Zi}	Every component of an input vector is subjected to the usual exponential function. It results in a positive number greater than zero, which will be extremely little if the input is negative and huge if the input is vast. It is not set within range (0, 1), which is what a probability must be.
$\sum_{j=1}^{K} e^{Zi}$	The normalization term is the term in the formula that appears at the bottom. It guarantees that the function's output values will add up to 1 and fall inside the range (0, 1), forming a legitimate probability distribution.
K	how many classes the multi-class classifier can handle.

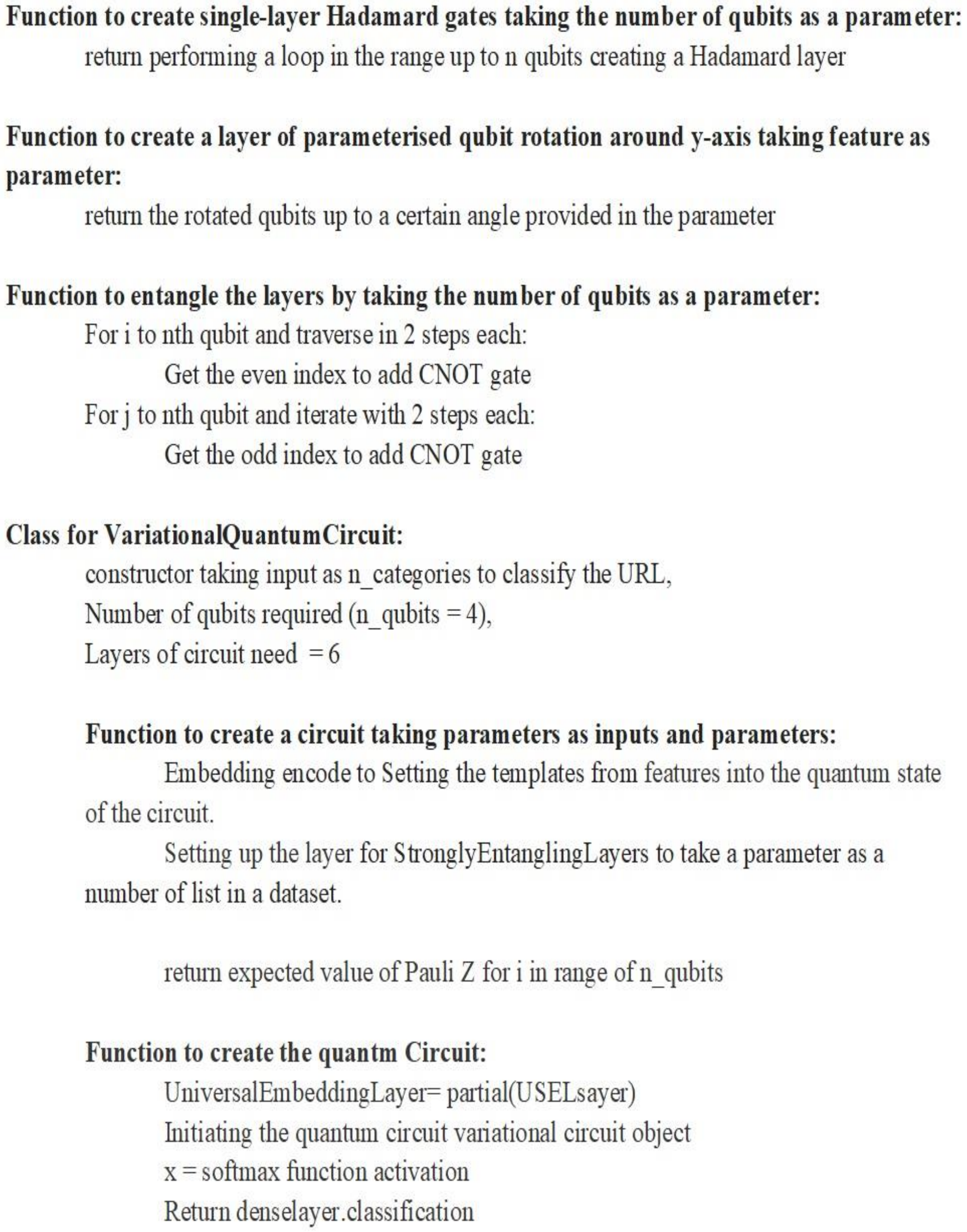
To "quantize" the proposed model, the authors substituted a variational quantum circuit for the layer between the embeddings and the output. A traditional dense layer typically has N inputs and M outputs; therefore, internally, it corresponds to matrix multiplication followed by bias addition and application of the activation function. However, quantum layers cannot accomplish this openly. Therefore, it implies that a quantum variational layer comprises three processes.
A traditional dense layer converts N inputs to N qubits and scales the input by π/2. (so it can represent a rotation around the Bloch sphere).

The Pennylane library accomplishes the three processes. To do this with the PennyLane library, one must first define a device that will perform the quantum operations. The IBMQ or Rigetti provides the real devices . Then, the python function encodes the actual circuit.
The PennyLane Library is a cross-platform open-source software development kit for

differentiable quantum computer programming. Classical calculations, such as model optimization or training, are carried out using typical scientific computing or machine learning libraries, such as SciPy in Python. PennyLane interfaces with these libraries and integrates them with quantum simulators.

Many dense layers may be stacked on top of each other to enhance the depth of a network; quantum variational layers can do the same.

Figure 3 depicts the pseudocode of implemented quantum encoding.



```
Function to create single-layer Hadamard gates taking the number of qubits as a parameter:
        return performing a loop in the range up to n qubits creating a Hadamard layer

Function to create a layer of parameterised qubit rotation around y-axis taking feature as
parameter:
        return the rotated qubits up to a certain angle provided in the parameter

Function to entangle the layers by taking the number of qubits as a parameter:
        For i to nth qubit and traverse in 2 steps each:
                Get the even index to add CNOT gate
        For j to nth qubit and iterate with 2 steps each:
                Get the odd index to add CNOT gate

Class for VariationalQuantumCircuit:
        constructor taking input as n_categories to classify the URL,
        Number of qubits required (n_qubits = 4),
        Layers of circuit need  = 6

        Function to create a circuit taking parameters as inputs and parameters:
                Embedding encode to Setting the templates from features into the quantum state
        of the circuit.
                Setting up the layer for StronglyEntanglingLayers to take a parameter as a
        number of list in a dataset.

                return expected value of Pauli Z for i in range of n_qubits

        Function to create the quantm Circuit:
                UniversalEmbeddingLayer= partial(USELsayer)
                Initiating the quantum circuit variational circuit object
                x = softmax function activation
                Return denselayer.classification
```

Fig 3 Pseudocode of Quantum Encoding

Further discussion covers Quantum Classification Variational Circuit. The execution of a quantum circuit requires a variety of complex pre- and post-processing steps. Figure 4 shows the quantum classification variational circuit.

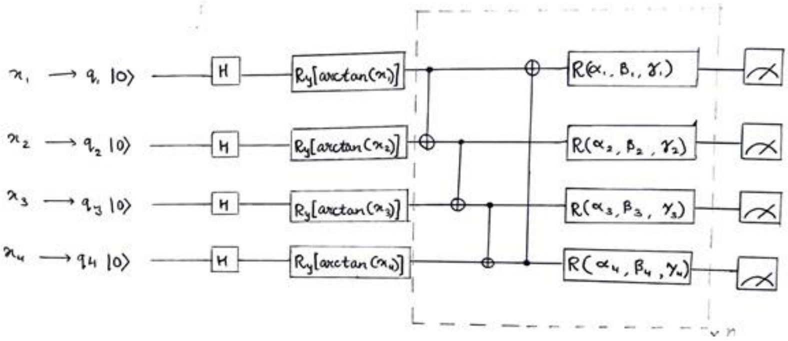

$x_1 \rightarrow q_1 |0\rangle$
$x_2 \rightarrow q_2 |0\rangle$
$x_3 \rightarrow q_3 |0\rangle$
$x_4 \rightarrow q_4 |0\rangle$
H
H
H
H
$R_y[\arctan(x_1)]$
$R_y[\arctan(x_2)]$
$R_y[\arctan(x_3)]$
$R_y[\arctan(x_4)]$
$R(\alpha_1, \beta_1, \gamma_1)$
$R(\alpha_2, \beta_2, \gamma_2)$
$R(\alpha_3, \beta_3, \gamma_3)$
$R(\alpha_4, \beta_4, \gamma_4)$
$\times n$

Fig 4 Quantum Classification Variational Circuit

An impartial beginning state is created using the Hadamard gate H. The three sections of the Variational Quantum Circuit (VQC) employed in this study: the encoding portion, the variational part containing parameters to be learned, and the measurement part, which will provide the Pauli-Z expectation values through repeated quantum circuit runs.

The first k qubits would be used for the quantum measurement. Arctan (x) is used for state preparation and is taken straight from the classical input data (cleaned words scrapped from the HTML of onion services). x1, x2, x3, and x4 represent the encoding of classical data into quantum data by Universal Sentence Encoder. All the qubits are at their Zeroth State before starting the circuit.

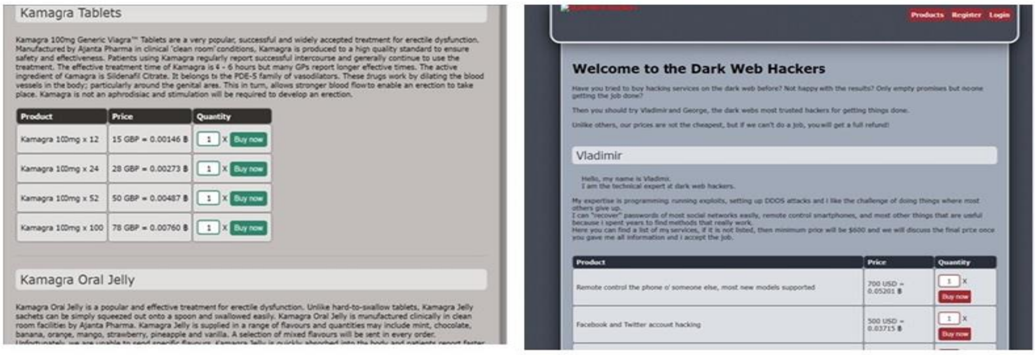
The number of classes is k. The VQC's expressivity is boosted by iterating part of the circuit highlighted by dashed lines.

The number of Qubits chosen depends on the input dimension of the data and the depth of circuit one needs (e.g. 512 dimensions in implemented approach)

4. RESULT AND DISCUSSION

The authors designed a dark web crawler to collect onion services. The crawler ran at depth 3 with multi-threading mode. The crawler collected HTML pages of 19,000 version 2 onion services. In addition, the crawler collected the URL of the onion service and the title and metadata of the respective service. The authors excluded hidden services based on the criteria of non-English onion pages and onion pages of the Facebook deep web. Depth 3 crawling resulted in many Facebook hidden pages categorized as social networking sites. More number of Facebook pages skewed the dataset. Thus authors removed Facebook onion pages collected at depth 2 and depth 3. After preprocessing and Facebook post removal, the authors considered 2000 onion services for analysis.

Figure 5 shows samples of onion service pages.



Kamagra Tablets

Kamagra 100mg Generic Viagra™ Tablets are a very popular, successful and widely accepted treatment for erectile dysfunction. Manufactured by Ajanta Pharma in clinical 'clean room' conditions. Kamagra is produced to a high quality standard to ensure safety and effectiveness. Patients using Kamagra regularly report successful intercourse and generally continue to use the treatment. The effective treatment time of Kamagra is 4 - 6 hours but many GPs report longer effective times. The active ingredient of Kamagra is Sildenafil Citrate. It belongs to the PDE-5 family of vasodilators. These drugs work by dilating the blood vessels in the body; particularly around the genital area. This in turn, allows stronger blood flow to enable an erection to take place. Kamagra is not an aphrodisiac and stimulation will be required to develop an erection.

Product	Price	Quantity
Kamagra 100mg x 12	15 GBP = 0.00146 ฿	1 X Buy now
Kamagra 100mg x 24	28 GBP = 0.00273 ฿	1 X Buy now
Kamagra 100mg x 52	50 GBP = 0.00487 ฿	1 X Buy now
Kamagra 100mg x 100	78 GBP = 0.00760 ฿	1 X Buy now

Kamagra Oral Jelly

Kamagra Oral Jelly is a popular and effective treatment for erectile dysfunction. Unlike hard-to-swallow tablets, Kamagra Jelly sachets can be simply squeezed out onto a spoon and swallowed easily. Kamagra Oral Jelly is manufactured clinically in clean room facilities by Ajanta Pharma. Kamagra Jelly is supplied in a range of flavours and quantities may include mint, chocolate, banana, orange, mango, strawberry, pineapple and vanilla. A selection of mixed flavours will be sent in every order. Unfortunately, we are unable to send specific flavours. Kamagra Jelly is quickly absorbed into the body and patients report faster

(a) Sample Onion Service 1

Products Register Login

Welcome to the Dark Web Hackers

Have you tried to buy hacking services on the dark web before? Not happy with the results? Only empty promises but no-one getting the job done?

Then you should try Vladimir and George, the dark webs most trusted hackers for getting things done.

Unlike others, our prices are not the cheapest, but if we can't do a job, you will get a full refund!

Vladimir

Hello, my name is Vladimir.
I am the technical expert at dark web hackers.

My expertise is programming, running exploits, setting up DDOS attacks and I like the challenge of doing things where most others give up.
I can "recover" passwords of most social networks easily, remote control smartphones, and most other things that are useful because I spent years to find methods that really work.
Here you can find a list of my services, if it is not listed, then minimum price will be $600 and we will discuss the final price once you gave me all information and I accept the job.

Product	Price	Quantity
Remote control the phone o' someone else, most new models supported	700 USD = 0.05201 ฿	1 X Buy now
Facebook and Twitter account hacking	500 USD = 0.03715 ฿	1 X Buy now

(b)Sample Onion Service 2

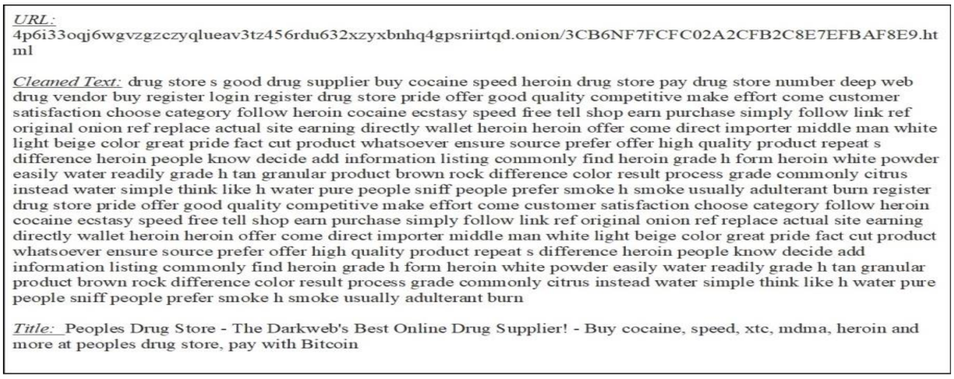
Fig 5 Sample Onion Services

Table 3 shows dataset samples with data attributes considered for quantum encoding.

Table 3 Dataset samples with data attribute considered for quantum encoding

Id	Onion Service	Cleaned_Text Data	Onion Service Title
1	4p6i33oqj6wgvzgzczyqlueav3tz456rdu632xzyxbnhq4gpsriirtqd.onion/3CB6NF7FCFC02A2CFB2C8E7EFBAF8E9.html	visit view visit use experience people completely new way b c d e f g h j k l m n o p q r s t u v w x y z check united kingdom grad visit united kingdom new More…	Peoples Drug Store - The Darkweb's Best Online Drug Supplier! - Buy cocaine, speed, xtc, mdma, heroin and more at peoples drug store, pay with Bitcoin
2	canxzwmfihdnn7bz.onion/A7045E8B1F9E04E6GF35DF8b96C29.html	chat accessible new shiny v3 address notbumpz34bgbz4yfdigxvd6vzwtxc3zpt5imukgl6bvip2nikdmdaad onionnotbumpz34bgbz4yfdigxvd6vzwtxc3zpt5imukgl6bvip2nikdmdaad onion chat group chat More…	Ableonion
3	bepig5bcjdhtlwpgeh3w42hfftcqmg7b77vzu7ponty52kiey5ec4ad.onion\F7E89E9EDFFByAKxE8c3E99C6s.html	quality original cheap paymentkamagra4bitcoin buy cheap ship login register login register mg generic popular successful widely accept treatment clinical clean room produce high quality standard ensure safety effectiveness More…	Kamagra For Bitcoin - Same quality as original viagra pills, cheap prices, Bitcoin payment

Figure 6 shows a snapshot of sample cleaned text for one onion service.



URL:
4p6i33oqj6wgvzgzczyqlueav3tz456rdu632xzyxbnhq4gpsriirtqd.onion/3CB6NF7FCFC02A2CFB2C8E7EFBAF8E9.html

Cleaned Text: drug store s good drug supplier buy cocaine speed heroin drug store pay drug store number deep web drug vendor buy register login register drug store pride offer good quality competitive make effort come customer satisfaction choose category follow heroin cocaine ecstasy speed free tell shop earn purchase simply follow link ref original onion ref replace actual site earning directly wallet heroin heroin offer come direct importer middle man white light beige color great pride fact cut product whatsoever ensure source prefer offer high quality product repeat s difference heroin people know decide add information listing commonly find heroin grade h form heroin white powder easily water readily grade h tan granular product brown rock difference color result process grade commonly citrus instead water simple think like h water pure people sniff people prefer smoke h smoke usually adulterant burn register drug store pride offer good quality competitive make effort come customer satisfaction choose category follow heroin cocaine ecstasy speed free tell shop earn purchase simply follow link ref original onion ref replace actual site earning directly wallet heroin heroin offer come direct importer middle man white light beige color great pride fact cut product whatsoever ensure source prefer offer high quality product repeat s difference heroin people know decide add information listing commonly find heroin grade h form heroin white powder easily water readily grade h tan granular product brown rock difference color result process grade commonly citrus instead water simple think like h water pure people sniff people prefer smoke h smoke usually adulterant burn

Title: Peoples Drug Store - The Darkweb's Best Online Drug Supplier! - Buy cocaine, speed, xtc, mdma, heroin and more at peoples drug store, pay with Bitcoin

Fig 6 Snapshot of sample cleaned text of onion service

The quantum encoding model received 2000 hidden service samples and achieved 99.6 % accuracy.
Table 4 shows the quantum model's sampled output label of the onion service categorization.

Table 4 Categorization of the onion services with quantum encoding

Id	Onion Service	Title	Category_ Label
1	4p6i33oqj6wgvzgzczyqlueav3tz456rdu632xzyxbnhq4gpsriirtqd.onion/3CB6NF7FCFC02A2CFB2C8E7EFBAF8E9.html	Peoples Drug Store-The Darkweb's Best Online Drug Supplier! - Buy cocaine, speed, xtc, mdma, heroin and more at peoples drug store, pay with Bitcoin	E-Commerce
2	ctemplarpizuduxk3fkwrieizstx33kg5chlvrh37nz73pv5smsvl6ad.onion\0F2AE691K5B841915B81BED0743CEA.html	The Only Anonymous Payment Resources You Will Ever Need? - CTemplar	Business/Corporate
3	bepig5bcjdhtlwpgeh3w42hffftcqmg7b77vzu7ponty52kiey5ec4ad.onion\F7E89E9EDFFByAKxE8c3E99C6s.html	Kamagra For Bitcoin - Same quality as original viagra pills, cheap prices, Bitcoin payment	Health & Fitness
4	ar.facebookcorewwwi.onion\6ADH89JBD10FB2A1B1084DABFF424.html	Belarus Solidarity Foundation	Law & Government

Figure 7 shows a pie chart representation of the categorization of different onion services

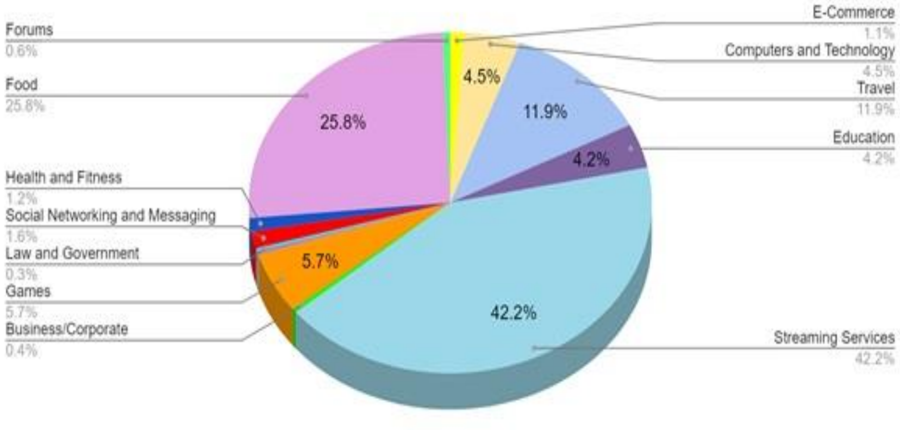
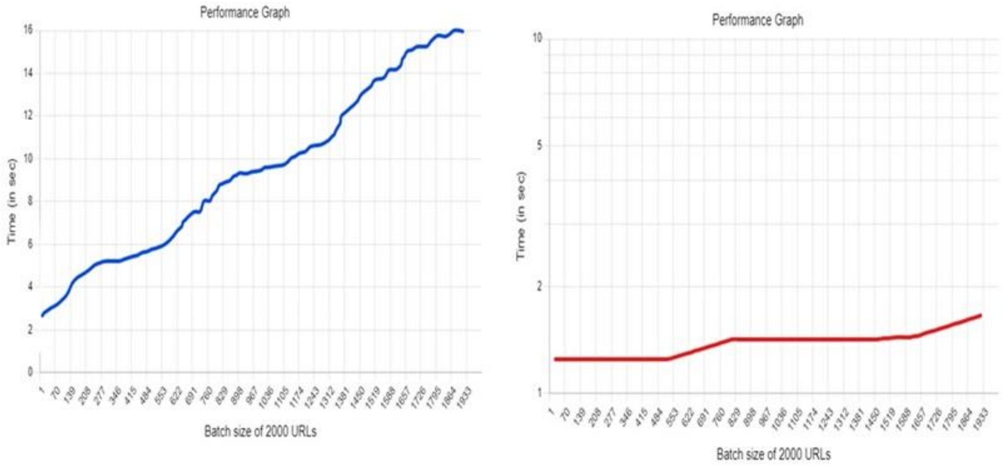


Fig 7 Categorization of different onion services

The authors investigated performance improvement with quantum encoding. Dark web crawler crawled was not focused on a particular domain; it crawled data continuously. Therefore data collection was huge. The proposed study aims to categorize dark web onion services with better performance by converting classical data into quantum encoding.

Time and Memory performance evaluation of Classical v/s Quantum model

Figure 8 depicts the comparison between classical and quantum models. In classical mode, with an increase in batch size, and time, consumption increases linearly. In quantum encoding, if the word size in a batch varies, if reflected with a slight variation in time consumption.



(a) Classical Model

(b) Quantum Model

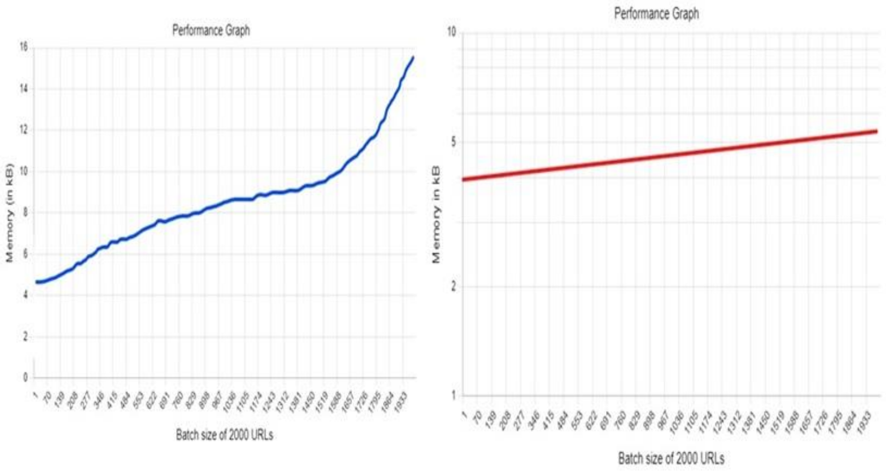
Fig 8 Time Required Vs Batch Size of 2000 Onion services

The cleaned text of onion services will be of different sizes. For example, onion service A has 5000 words, whereas onion service B has 7500 words. The classical model takes comparatively more time to categorize the onion service B because of the content-based classification approach. Thus, as the batch size increase, the overall categorization time increases.

In quantum encoding, the incoming cleaned texts from the classical model were converted to quantum data. Although quantum data has less memory as qubits and qubits can represent an exponential number of bits. The overall time slightly increases when sample text has more than 8000 words that cannot represent the whole data in qubits. Model discarded the words that did not fit in a single interaction (processing a single onion service). Also, before the final output, the data is passed through the Softmax function to ensure the output value is correct/matched.
Another performance constraint with dark web data categorization is memory consumption.

Figure 9 shows the comparison between classical and quantum models for memory consumption. In the classical model, the incoming cleaned texts of different onion services have different sizes. For example, onion service A has 5000 words, and onion service B has 7500 words. Thus, the memory consumption by the model will increase.

In the quantum model, the qubits are conserved in every iteration (classifying the one onion service at a time). It means that before and after the output, the number of qubits involved in the circuit is equal. The value of less significant words in classification stays in the qubit until the following words are encoded into the same qubit.



(a) Classical Model

(b) Quantum Model

Fig 9 Memory Required Vs Batch size of 2000 URLs

5. CONCLUSION

Dark web data analysis is proven one of the proactive cyber security measures; therefore, researchers attempt to collect and label dark web data. Such data could help draw meaningful inferences for cyber security. In the present work, the authors developed a crawler to crawl the surface and dark web. The crawler collected a substantial amount of data from all webs. As a result, traditional data analysis and processing machine learning still require a computationally intensive and time-consuming process.

Therefore, the authors attempted to apply the concept of quantum encoding—the classical data fed to the quantum encoding circuit. Input hidden services were categorized using a quantum encoding circuit. The softmax function is applied before the display category of hidden service.

The proposed work utilized a pre-trained version of Google's Universal Sentence Encoder (USE) model. The model performed well and delivered accurate outcomes for a range of transfer tasks. The authors discussed the relationship between model complexity, resource utilization, the accessibility of training data for transfer tasks, and performance outcomes for both models. Pre-trained word embedding baselines with word-level transfer learning model contrasted against baselines without transfer learning. Work observed that sentence-level transfer learning performs better than word-level transfer. Transfer learning with word embeddings surpasses sentence-level transfer, we find. The proposed work shows good performance with small amounts of supervised training data for a transfer task employing transfer learning via word embeddings.

Further, the result concludes that the quantum encoded model required much less memory and performed computations much faster on data volume than the classical machine learning model. Thus quantum computing enabled model could be an efficient choice for designing a dark web data analysis framework.

The representation must be small to use modern NISQ devices and employ only a few qubits and quantum gates. Qubits decay quickly, and quantum gates are also error-prone, restricting the number of operations required to establish the quantum state, which must be modest.
Quantum classification versions take significantly longer to train using the simulator (approximately one minute per epoch with 32 batches). It is possible that quantum simulation is time-consuming and becomes increasingly complex as the number of qubits increases.

Although the USE embedding facilitated categorization tasks in the proposed work, the future scope of work could use transfer learning instead of having to train a model.

REFERENCES

1. Craig, A. N., Shackelford, S. J., & Hiller, J. S. (2015). Proactive cybersecurity: A comparative industry and regulatory analysis. American Business Law Journal, 52(4), 721-787.
2. Meland, P. H., Tokas, S., Erdogan, G., Bernsmed, K., & Omerovic, A. (2021). A Systematic Mapping Study on Cyber Security Indicator Data. Electronics, 10(9), 1092.
3. Arnold, N., Ebrahimi, M., Zhang, N., Lazarine, B., Patton, M., Chen, H., & Samtani, S. (2019, July). Dark-net ecosystem cyber-threat intelligence (CTI) tool. In *2019 IEEE International Conference on Intelligence and Security Informatics (ISI)* (pp. 92-97). IEEE.
4. Shakarian, P. (2018). Dark-web cyber threat intelligence: from data to intelligence to prediction. Information, 9(12), 305.
5. Zhang, X., & Chow, K. P. (2020). A framework for dark Web threat intelligence analysis. In Cyber Warfare and Terrorism: Concepts, Methodologies, Tools, and Applications (pp. 266-276). IGI Global.
6. Samtani, S., Zhu, H., & Chen, H. (2020). Proactively identifying emerging hacker threats from the dark web: A diachronic graph embedding framework (d-gef). ACM Transactions on Privacy and Security (TOPS), 23(4), 1-33.
7. Jeziorowski, S., Ismail, M., & Siraj, A. (2020, March). Towards image-based dark vendor profiling: an analysis of image metadata and image hashing in dark web marketplaces. In Proceedings of the Sixth International Workshop on Security and Privacy Analytics (pp. 15-22).
8. Meland, P. H., Bayoumy, Y. F. F., & Sindre, G. (2020). The Ransomware-as-a-Service economy within the darknet. Computers & Security, 92, 101762.
9. Koloveas, P., Chantzios, T., Tryfonopoulos, C., & Skiadopoulos, S. (2019, July). A crawler architecture for harvesting the clear, social, and dark web for IoT-related cyber-threat intelligence. In 2019 IEEE World Congress on Services (SERVICES) (Vol. 2642, pp. 3-8). IEEE.
10. Al Nabki, M. W., Fidalgo, E., Alegre, E., & De Paz, I. (2017, April). Classifying illegal activities on tor network based on web textual contents. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers (pp. 35-43).
11. Guitton, C. (2013). A review of the available content on Tor hidden services: The case against further development. Computers in Human Behavior, 29(6), 2805-2815.
12. Faizan, M., Khan, R. A., & Agrawal, A. (2020). Ranking potentially harmful Tor hidden services: Illicit drugs perspective. Applied Computing and Informatics.
13. Huete Trujillo, D. L., & Ruiz-Martínez, A. (2021). Tor Hidden Services: a systematic literature review. Journal of Cybersecurity and Privacy, 1(3), 496-518.
14. Zhao, F., Zhou, J., Nie, C., Huang, H., & Jin, H. (2015). SmartCrawler: a two-stage crawler for efficiently harvesting deep-web interfaces. IEEE transactions on services computing, 9(4), 608-620.
15. Frank, R., Westlake, B., & Bouchard, M. (2010, July). The structure and content of online child exploitation networks. In ACM SIGKDD Workshop on Intelligence and Security Informatics (pp. 1-9).
16. Zulkarnine, A. T., Frank, R., Monk, B., Mitchell, J., & Davies, G. (2016, September). Surfacing collaborated networks in dark web to find illicit and criminal content. In 2016 IEEE Conference on Intelligence and Security Informatics (ISI) (pp. 109-114). IEEE.
17. Dalvi, A., Paranjpe, S., Amale, R., Kurumkar, S., Kazi, F., & Bhirud, S. G. (2021, May). SpyDark: Surface and Dark Web Crawler. In 2021 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC) (pp. 45-49). IEEE.
18. Dong, F., Yuan, S., Ou, H., & Liu, L. (2018, November). New cyber threat discovery from darknet marketplaces. In 2018 IEEE Conference on Big Data and Analytics (ICBDA) (pp. 62-67). IEEE.
19. Dalins, J., Wilson, C., & Carman, M. (2018). Criminal motivation on the dark web: A categorization model for law enforcement. Digital Investigation, 24, 62-71.
20. He, S., He, Y., & Li, M. (2019, March). Classification of illegal activities on the dark web. In Proceedings of the 2019 2nd International Conference on Information Science and Systems (pp. 73-78).

21. Mahor, V., Rawat, R., Kumar, A., Chouhan, M., Shaw, R. N., & Ghosh, A. (2021, September). Cyber warfare threat categorization on cps by dark web terrorist. In 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON) (pp. 1-6). IEEE.
22. Takaaki, S., & Atsuo, I. (2019, March). Dark web content analysis and visualization. In Proceedings of the ACM International Workshop on Security and Privacy Analytics (pp. 53-59).
23. Kobayashi, H., Kadoguchi, M., Hayashi, S., Otsuka, A., & Hashimoto, M. (2020, November). An expert system for classifying harmful content on the dark web. In 2020 IEEE International Conference on Intelligence and Security Informatics (ISI) (pp. 1-6). IEEE.
24. Kinder, A., Choo, K. K. R., & Le-Khac, N. A. (2020). Towards an Automated Process to Categorize Tor's Hidden Services. In Cyber and Digital Forensic Investigations (pp. 221-246). Springer, Cham.
25. Ghosh, S., Das, A., Porras, P., Yegneswaran, V., & Gehani, A. (2017, August). Automated categorization of onion sites for analyzing the darkweb ecosystem. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1793-1802).
26. Buldin, I. D., & Ivanov, N. S. (2020, January). Text classification of illegal activities on onion sites. In 2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus) (pp. 245-247). IEEE.
27. Moraliyage, H., Sumanasena, V., De Silva, D., Nawaratne, R., Sun, L., & Alahakoon, D. (2022). Multimodal Classification of Onion Services for Proactive Cyber Threat Intelligence using Explainable Deep Learning. IEEE Access.
28. Njorbuenwu, M., Swar, B., & Zavarsky, P. (2019, June). A survey on the impacts of quantum computers on information security. In 2019 2nd International conference on data intelligence and security (ICDIS) (pp. 212-218). IEEE.
29. Laxminarayana, N., Mishra, N., Tiwari, P., Garg, S., Behera, B. K., & Farouk, A. (2022). Quantum-Assisted Activation for Supervised Learning in Healthcare-based Intrusion Detection Systems. IEEE Transactions on Artificial Intelligence.
30. Ko, K. K., & Jung, E. S. (2021). Development of cybersecurity technology and algorithm based on quantum computing. Applied Sciences, 11(19), 9085.
31. Tosh, D., Galindo, O., Kreinovich, V., & Kosheleva, O. (2020, June). Towards security of cyber-physical systems using quantum computing algorithms. In 2020 IEEE 15th International Conference of System of Systems Engineering (SoSE) (pp. 313-320). IEEE.
32. Ali, A. (2021, January). A Pragmatic Analysis of Pre-and Post-Quantum Cyber Security Scenarios. In 2021 International Bhurban Conference on Applied Sciences and Technologies (IBCAST) (pp. 686-692). IEEE.
33. Suryotrisongko, H., & Musashi, Y. (2022). Evaluating hybrid quantum-classical deep learning for cybersecurity botnet DGA detection. Procedia Computer Science, 197, 223-229.
34. Abohashima, Z., Elhosen, M., Houssein, E. H., & Mohamed, W. M. (2020). Classification with quantum machine learning: A survey. arXiv preprint arXiv:2006.12270.
35. Kilber, N., Kaestle, D. and Wagner, S., 2021. Cybersecurity for Quantum Computing. arXiv preprint arXiv:2110.14701.