UNIVERSITY
OF TWENTE.

Investigating whether typosquatting targets children

Akshay Prasad - s2433583
University of Twente
PO Box 217, 7500 AE Enschede
the Netherlands
akshay.prasad@student.utwente.nl

Abstract—**Typosquatting refers to registering the domain names that are typo variants of the original domain. The study investigates whether typosquatting targets children by understanding some of the significant reading, writing, and typing errors they make and the domains they regularly use of various ages. We identified some popular typosquatting tools and assessed what percentage of children's errors are covered by the tool. To obtain the concrete evidence, we are checking the evidence of blacklisting from the DNSTwist with and without applying children's error categories, i.e. *Addition, Omission, and Substitution*. We performed the measurement continuously for about 30 days to see the stability in terms of the results. Once we determine the results, we compare them against the Alexa top records and conclude.**

Keywords **- Children reading, writing, and typing errors, Error categories, Domains, Typosquatting, DNSTwist, VirusTotal, Blacklisting, Measurement**

I. Introduction

Learning and writing a new language is typically complex for any individual in the beginning stages. In the same way, it is hard for the children who are learning to read and write to take into account characters, numbers, letters, punctuation, and articulations [1]. Thus, children of various ages may be prone to reading and writing errors. However, handwriting and typewriting(typing) are similar perceptual skills, and reading errors also lead to typing errors as well [2].

Typosquatting is the practice of registering a domain name with an intentionally misspelled version of the original brand or domain [3]. For example, 'www.google.com' can be registered as 'goo0gle.com'. The goal of the typosquatting is to redirect the users to unintended destinations or to steal the user traffic for various reasons such as monetary gains, et cetera [4]. Although typosquatting and children's reading and writing errors are part of two different aspects of society, we wonder if there may be an overlap with the similar traits of misspelled characters. For example, the letters "D" and "B" visually look similar with certain writing representation. Thus, there might be a chance that children substitute one of the characters in place of the other. Therefore, it is possible that children might end up using malicious websites.

Although there are millions of improper domains on the internet, some of the worst are just a mistake away, ready to happen. While recent studies in the United States of America report that 74% of children aged 8-18 years have access to the internet, in the Netherlands, practically all children are online nowadays. Usually, Dutch children use Google[1] as their primary source of information seeking [5]. Thus, the fact that, children could be potential victims of typosquatting is a concrete possibility.

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

A. Research questions

In this research, we investigate whether children are the target of typosquatting. This research involves understanding the reading and writing errors of children of various ages. Once the error categories of children of different ages are obtained, the common domains of children are listed based on the methodology decided, which will be discussed in the methodology section later. Once the domains are listed, it is assessed against the typosquatting tools to determine whether they cover the errors from children or not. The domains are further investigated to determine whether they are flagged as malicious. If they are flagged as malicious, we investigate if the blacklisted domains are part of the children error category. Also, we investigate if there are any changes in the blacklisting number over the period by doing the continuous measurement. Thus, the entire stated flow is fragmented into one main question and four sub-questions.

The research question and sub-questions are as follows:

The terminologies in the questions, RQ and SQ refers to Research Question and Sub-question, respectively.

RQ: Are children the target of Typosquatting?

- **SQ1:** What typical reading and writing errors do children of various ages generally make?
- **SQ2:** What are some of the most common websites or domains children regularly use of various age groups?
- **SQ3:** Do any of the existing typosquatting tools cover the type of errors that children make?
- **SQ4:** Do we see evidence of children's domains in the blacklisted list?

B. Technical contribution

This research provides significant techical contributions, such as semi-automation of the flow which are represented in the form of research questions which were mentioned in the earlier sections research questions I-A. The algorithm is developed to differentiate the visually similar letters or characters through it's unicode, a unique numerical form for a character, representation for a variety of domains obtained from the typosquatting tool.

This document will provide a literature review of the reading and writing errors children usually make, typosquatting and its several types, and existing tools that help list typosquatting domains. This research will investigate whether any existing typosquatting techniques cover children's errors. Also, this research will further investigate whether any of the children's domains are flagged as malicious. Section II describes the literature review, and Section III describes the methodology to solve the problem. Section IV describes the results obtained, followed by the conclusion and discussions with limitations in sections V and VI.

II. Literature Review

A. Cause of reading and writing errors in Children

Researchers have observed that children experience difficulties while reading and writing due to several factors. Some of them as follows:

- **Motor skills**
  Motor skills is commonly known as" touch-typing" when children learn to type in a formal setting. People should text with numerous fingers without glancing at the keyboard. Touch typing can be a tremendous challenge for children with fine motor skills. The coordination of small muscles in movement with the eyes, usually including the synchronization of hands and fingers, is a fine motor skill. It can be difficult to move one finger while using both hands on the keyboard to isolate a letter (or their thumb for the space bar). Using the trackpad or mouse adds to the difficulty.

- **Spatial challenges**
  Learning to type may also be complicated by spatial challenges. Children must hover over the middle ("home") row with one hand on the right side and the other on the left. Visual-spatial processing issues can make finding a specific letter amid a sea of keys

challenging. Children may also struggle to comprehend the spatial distance between the letters.

- **Memory issues**
Some children have difficulty remembering where the letters on the keyboard are located. As a result, they rely on their vision to find the correct letter, which slows them considerably. Also, children may rely on memory when picking phonetically similar words and are prone to making mistakes.

- **ADHD(Attention-Deficit/Hyperactivity Disorder) issues and executive functioning problems**
In typing, some children have difficulty focusing. As a result of focusing issue, children make errors. Too many ADHD children fall behind when reading, skipping words or sentences, losing sight of where they are on the page, and missing details and connections. These issues are especially noticeable in long and challenging sections.

- **Errors because of phonetically similar words**
The pronunciation of the phonetically similar words is identical to some degree, making it difficult for children to grasp the word's letters. For example, "*current see*" and "*currency*" are heard in identical ways but are written differently. Thus, children are prone to make errors here and discussed in detail in subsection B of the main section II.

- **Visually similar words**
In any language system, some words are visually similar. The letters are probably jumbled and form another word. Indeed, they are confusing for children if they rely on memorizing words. For example, "big" and "dig" is one case. There is specific literature based on the experiments conducted on children, described in subsection B of central section II.

B. Existing research on reading and writing mistakes from children

The cited paper by Mark, Shankweiler, Liberman, and Fowler [6] carried the experiment on phonetic recording and reading difficulty in beginning readers with 34 participants, where 18 children were good readers, and the rest were not. The average grade level of good readers was 3.97, ranging from 3.1 to 4.5, and the average grade level of wrong readers was 2.19, ranging from 1.5 to 2.4. The good readers had a mean age of 92.4 months, and the mean age of the poor readers was 94.0 months. From this experiment, it is drawn that good readers generally make more errors in rhyming words, and on the other hand, poor readers make typical errors in rhyming and non-rhyming words.

Similarly, from the citation of Swearingen [7], it is stated that children may spell words correctly, but when written, they are prone to making mistakes and vice versa. Through a simple spelling work experiment, Swearingen [7] analyzed third, fourth, fifth, and sixth-grade children to understand the kinds of spelling mistakes they make. The paper's analysis is to get information such as similarities and differences in mistakes when children spell from the lists, purposeful writing of copying the sentences, the frequency of errors while writing, et cetera. The nature of the errors seen during the experiment is as follows:

- **Many errors accounted for omissions** (apostrophes, capitals, silent letters such as *cach* for *catch* or *suden* for *sudden*).
- **Many accounted for Substitutions, usually phonetic** (*rane* for *rain*, *kar* for *care*, *plas* for the *place*, *erlee* for *early*, *ourwer* for an *hour*, *apon* for *upon*).

The several categories of reading and writing errors that children displayed during the experiment are tabulated in the table I.

TABLE I
TYPE OF READING AND WRITING ERRORS

Type of Error	Grade 3	Grade 6
Omissions, substitutions	37%	48%
Including phonetic substitutions	46%	43%
Reversal	7%	3%
Addition	7%	3%
Incorrect pronunciation	3%	—

Thus, the errors were closely concentrated in the omission and substitution categories, such as list spelling and writing. Many of the substitutions were applications of related phonetic knowledge [7].

Also, the cited article by Hughes and Wilkins [8] investigated the reading schemes of various children ages 5 to 11 years. They involved 120 children reading four passages that adopted the typography of four reading stages. The spacing of the texts decreased with each stage as the reading age increased. It is inferred from the experiment that the reading speed decreased for the children aged between 5 to 7 years old as the font or the text size decreased. Children aged between 8 and 11 years old did not have many disadvantages due to the text size. However, children of all ages, mainly those susceptible to visual stress, made errors on the more minor texts compared to the larger ones.

As stated in the article by the authors Walker and Reynolds [9], there is a view that Sans serif font type are suitable for beginner readers, infants. In the paper it is also argued that, Serif typeface have greater and better variation in the thickness of the letter strokes than the sans serifs. Furthermore, some sans serif faces are built on a modular foundation, with numerous geometric shapes coupled in various ways to generate letterforms, whereas serif faces are typically based on the shapes of stone-cut or handwritten letterforms. However, in the paper by Wilkins et al[10], it is also mentioned that most typeface are sans serif for children. In addition, as stated in the paper by Simpson et al [11], letter confusion errors, or providing the name or sound of a letter other than the one offered (e.g., replying with the name of ”b” when presented with the letter ”d”), are one of the many sorts of errors that children make in such activities.

As stated in the citation, the author Read et al [12] compared the usable on text inputs such as mouse, keyboard, handwriting and speech recognition for the age group 6 to 8 years old. They concluded that the QWERTY keyboard layout adds extra pressure on children in terms of their user’s memory. Some of the type of errors and their examples are presented in the table II.

TABLE II
TYPE OF ERRORS OBSERVED

Type of Error	Example and Description
Cognition error	Child read a word wrong or cannot differentiate letters.
Spelling error	Child read a word wrong or pronounces a word wrong that they are aware off.
Selection error	Child will pick wrong letter, maybe ’I’ for ’i’.
Construction error	Child cannot form the letter or word correctly. In handwriting, ’a’ may look like ’d’. In speech, ’dragon’ becomes ’dwagon’.
Execution error	The child sometimes may fail to hit the adjacent character while typing or writing.
Software Induced Error	Software does wrong recognition of a word or character.

Also, the authors Read, and Horton [13], carried out an experiment with teenagers aged 13-14 years. The task is text copying to analyze typing errors. After experimenting, the authors classified the errors into several categories. The most important ones are Next-to-errors(NT), where the teenagers type the adjacent character instead of the intended character. Another error is Close Errors(CE), where the teenagers type the diagonally adjacent characters instead of intended characters. The other types of error which may be less effective for our context could be space errors within a single word where an unnecessary space is provided while writing a word.

On the other hand, the author White [14] created typing drills to address the typing errors. Experimenting with the typing task, the author analyzed approximately 20 thousand typing errors based on the QWERTY keyboard. The author categorized all these errors into several categories, which can be seen in the table listed below.

TABLE III
ERROR CATEGORIES FROM TYPING DRILLS

Type of Error	Percentage of total error
Substitution error	40%.
Omission error	20%.
Spacing	15%.
Transposition error	15%.
Insertion error	3%
Double typed error	2%
Capitalization error	2%

The categories of error can be seen from the table III and however, the definition for the above mentioned errors are not explicitly mentioned by the author but used the errors to design better typing tasks.

C. Typosquatting and the existing types of typosquatting

Typosquatting refers to registering domain names that are typo variants of popular domains. For example, users may mistype ”Google.com” as ”Gooogle.com”. Also, from the citation [15], typosquatting refers to registering domain names that are typos of their target domains, which usually host domains with significant traffic. In other words, the typosquatting domains that depict the original domains or the domains will attract users toward them for various reasons, capitalizing on their genuine typing mistakes. The individuals or organizations who register typosquatting domains (typo domains) are called ”typo squatters”. The figure 1 provides an example of typosquatting for the domain **utwente.nl**.

1) Typosquatting generation techniques: The first and widely cited approaches in the area of typosquatting generation techniques are given by Wang et al. [15]. The several types of generation models considering the example, www.utwente.nl, are listed below:

- **Missing-dot typos**
  The missing dot typing error occurs when the dot following “www” is forgotten in the written text, for example, ***wwwutwente.nl***.
- **Character-omission typos**
  The character-omission typing error occurs when one character in the original text is omitted, such as ***www.utwnte.nl***.
- **Character-permutation typos**
  The character-permutation typing error occurs when two consecutive characters are swapped in the original text and, for example, ***www.utwnete.nl***.
- **Character-substitution typos**
  The character-substitution typing error occurs when characters are replaced in the text by their adjacent ones on a specific keyboard layout, for example, ***www.utwemte.nl***, where ***“n”*** is replaced by the QWERTY-adjacent character ***“m”***.
- **Character-duplication typos**
  The character-duplication typing error occurs when typing characters accidentally twice, where they appear once in the original text, for example, ***www.utweente.nl***.

A similar approach was followed by Banerjee et al. [17] for generating typosquatted domains and are as follows:

- **1-mod-inplace**
  The 1-mod-inplace typing error occurs when the typosquatter substitutes one letter or character from the original domain name with all possible alphabet letters.
- **1-mod-deflate**
  The 1-mod-deflate typing error occurs when a typosquatter removes one letter or character from the original domain name or the domain.
- **1-mod-inflate**
  The 1-mod-inflate typing error occurs when a typosquatter increases the length of the domain by one.

D. Different typosquatting attacks

The typosquatting generation techniques were mentioned and discussed in the earlier sections II-C1, and more typosquatting techniques exist that exploits sound and visual similarities of the text. Such techniques and attacks based on that are as follows:

- **Homograph attack**
  Homograph attack relies on visually similar letters that might confuse users while writing or reading. For example, the letter ’l’ is

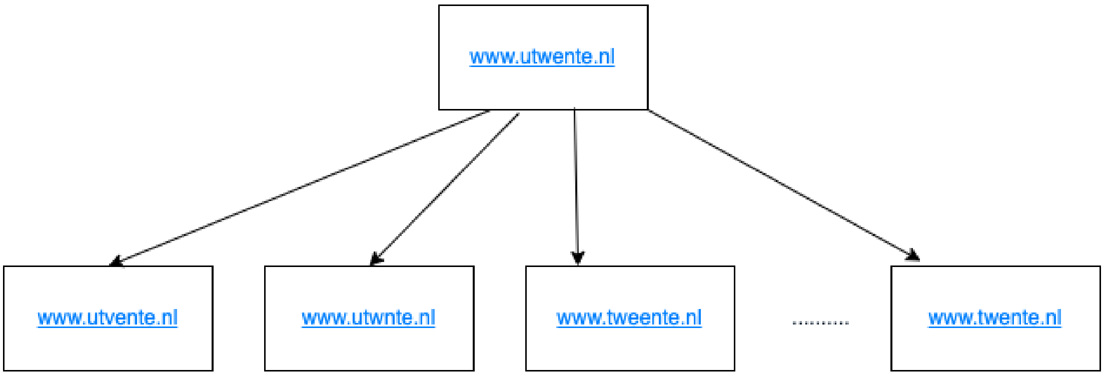


Fig. 1. TYPOSQUATTING DOMAINS FROM ORIGINAL BRAND (DIAGRAM CREATED ONLINE [16])

identical to‘ i’ with the font sans-serif. The article [2] suggests that this type of attack is not severe but a desirable choice for the attackers. For example, the typosquatting domain www.paypai.com targets the popular payment site PayPal, www.paypal.com.

A unique identifier represents every character in the real world Unicode or ASCII code. Despite the plethora of characters of several languages, some characters resemble or look visually similar and are called **Homoglyphs**. As stated in the cited paper [18], the Unicode system contains visually similar characters called homoglyphs. An attacker can register a domain that may not be visually recognizable using the homoglyph technique. Also, as cited by Max and Stuart [19], a homoglyph assault is a deception method that employs homoglyphs or homographs to build fake domains of existing brands to deceive consumers into clicking. A homoglyph is a combination of two or more symbols or glyphs with identical or extremely similar forms. For example, the Latin small letter ’O’(U+006F) and the Digit zero, ’0’ (U+0030) look visually similar. In addition, another example would be the Latin letter L ‘ı’ (U+0131) looks visually similar to exclamation mark ‘!’ (U+0021). Here, the ’U+006F’, ’U+0030’, ’U+0131’, and ’U+0021’ are unicode representations.

- **Soundsquatting attack**
  As per the article [3], the soundsquatting takes advantage of the similarity of words regarding sound and user confusion on which word fits the context. Soundsquatting, unlike typosquatting, is based on homophones, which are two words that sound the same but are spelt differently. For example, ”ate” and ”eight”.

- **Bitsquatting attack**
  An attacker leverages random bit-errors occurring in the memory of commodity computers and smartphones, to redirect internet traffic to attacker-controlled domains is called a bitsquatting attack [20]. From the fig 2 it can be seen that the single bit flip in a Q is resulting in S.

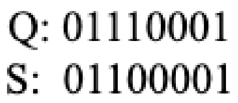

Q: 01110001
S: 01100001

Fig. 2. BITSQUATTING: THE RESULT OF SINGLE BIT FLIP

E. Existing tools to identify typosquatting attacks

- **DNSTwist**
  DNSTwist is a Python software that allows users to detect phishing, typo squatters, and attack domains based on an inputted domain, as described by [21]. Suppose the user is the owner of a domain or in charge of the company’s domains administration and brand protection. In that case, this tool can be quite valuable in identifying sites attempting to damage others by impersonating the user brands.

Through the technique of permutation this tool generates several possibilities of typosquatting such as,

- **Addition**
  As discussed in the section II, this typing error is due to adding an extra character in the original text.
- **Bitsquatting**
  As described in the section II of the literature review, this is because of flipping a bit in the original value.
- **Insertion**
  This type of typosquatting occurs when an extra letter is added to the original string.
- **Omission**
  This type of typosquatting occurs when the letter is removed from the original brand.
- **Repetition**
  This type of typosquatting occurs when one letter is repeated while typing.
- **Replacement**
  This type of typosquatting occurs when one letter is replaced by the other letter.
- **Homoglyph**
  As discussed in the earlier sections II-D, homoglyph is a combination of two or more symbols or glyphs that have identical or extremely similar forms. For example, the Latin small letter O (U+006F) and the Digit zero (U+0030) look visually similar.
- **Hyphenation**
  This is a form where the words are connected or separated through a hyphen. For example, the word ”son-in-law” is an example of hyphenation. On the other hand, it is misleading sometimes to add hyphen in the normal word and one such example is ”*cartoonnetwork*”, where one can hyphenate and type ”*cartoon-network*”.
- **Transposition**
  The transposition of a character is basically the adjacent characters are swapped or replaced. For example, for the word ”*form*”, it is possible to write ”*from*”.
- **vowel-swap**
  The vowel-swap is the replacement of a vowel character with another vowel character. For example, in the original string ”*cartoonnetwork.nl*” the vowel character is replaced by another vowel character and it is ”*cartoannetwork.nl*”.

- **Strider URL Tracer with Typo-patrol**
  The Strider URL or domain Tracer is a program that allows users to look at third-party domains that are called behind the scenes when a user visits a first-party domain. It has a top domains feature showing the most popular third-party domains. It also has a Typo-Patrol tool that builds and analyses typo-squatting domains for a particular target domain automatically [22].

- **Typodomain**
  The domain name represents the brand. Typodomain is a brand analysis and monitoring tool that reports typosquatters’ abuse towards the original brand. Typo domains are identical to the brand domain and are registered to capture visitors. Domain owners have the right to own and hold variations of valuable domains with themselves [23].

  Typo domains come in a few broad types, for example, youtube.com: Common misspellings and typing errors could be wwwyoutube.com or youtub.com, or youtube.pl, and such variations of typo domains are discussed in the earlier sections.

F. Blacklisted domains

Malicious content has proliferated along with the eruption of the internet. Therefore, many organizations build and maintain blacklists to help customers protect their computers [24]. Also, as mentioned in the paper by Velden [25], there are two types of blacklist vendors and are public domain blacklists and private domain blacklists. The public domain blacklists are available and can be accessed by anyone if they need them for their purpose. On the other hand, the private blacklists can be accessed quickly, requiring some subscription to obtain the list.

Some of the popular examples of public blacklists are VXVault [26], URLHaus [27] , and VirusTotal [28]. On the other hand, some famous examples of private blacklists are from antivirus vendors like Norton, McAfee and Spam Haus [29]. Also, as described in the article [25], the organizations usually maintain blacklisted domains as the domain may correspond to a phishing scam, malware, et cetera. A phishing scam is where the attacker sends a fraudulent message to a person and victimizes them into revealing sensitive information to the attacker. Malware is a computer program designed for computer disruption, leak sensitive data and also helps in gaining illegal entry into the system. However, we are using DNSTwist for this research to generate the typosquatting domains, as explained in the earlier sections II-E. To verify the legitimacy of the typosquatting domains that the DNSTwist generates, we use VirusTotal [28].

G. Text comparison with similarity metrics

The text similarity is to identify how similar the given two texts are. However, the same strategy of comparing two texts is also applied to comparing domains. The various algorithms available for the text comparison is discussed in the article [30] are as follows:

*1) **Levenshtein distance:*** The Levenshtein distance is the least number of single-character edits required to transform one word into the other, yielding a positive integer that is sensitive to string length. For example, the distance between *foo* and *bar* is 3. Firstly, the 'f' of the first word is different from the 'b' in the second word. Secondly, 'o' and 'a' are different, and thirdly, the 'o' and 'rare different. Thus it needs three edits to transform the word. Also, in other words, it is stated that the minimum number of addition, omission, and substitutions is needed to convert one word to another one. In the Levenshtein comparison metrics there are a plethora of variants and one them is:

Damerau-Levenshtein

The normal version of Levenshtein uses the strategy of the minimum number of addition, deletion, and substitutions to compare the words. Still, this variant adds the operation of transposition, which consists of swapping two characters.

*2) **Cosine distance:*** The cosine distance is used to measure the distance between two vectors. For cosine distance, the words are represented in vectors and can be compared by evaluating the angle between the two words of vectors. The cosine similarity results range from 0 to 1, where 0 is the least similarity and 1 is the greatest similarity. For example, the cosine similarity of the words *"netflix"* and *"neftlix"* is 0.8571.

*3) **Euclidean distance:*** The Euclidean distance is used to calculate the distance between two points. The consolidated formula is as follows: For the given point (x1,y1), and (x2, y2), the euclidean distance is calculated as follows:

$$\sqrt{(x1 - x2)^2 + (y1 - y2)^2}$$

For a small example, consider (x1,y1) as (2,-1) and (x2,y2) as (-2,2) and on the applying the mentioned formula earlier, the resulting value will be 5.

However, given the original string for the comparison, the Euclidean distance technique can help find the distance between the intended letter and the misspelled letter within the QWERTY keyboard layout. Thus, even though the technique cosine distance can be used to obtain the similarity between the words, in this research, we will be using Levenshtein along with its variant Damerau as its operations are similar to children's errors of Addition, Omission, and Substitutions. However, for some categories of addition and substitutions, such as adjacent key replacement in the written word, the Euclidean distance is used where 1 is the replacement of adjacent keys to the intended keys.

H. Phonetic similarity of words

The words are said to be phonetically similar if they sound similar during the pronunciation. However, for computers, some algorithms help determine the phonetic similarities between the words. Some of the algorithms are as follows:

- **Soundex, and Metaphone**
  As suggested in the paper by Koneru, Pulla, and Varol [31], the Soundex and Metaphone are quite popular phonetic matching algorithms. The Soundex algorithm is the first version

compared to the Metaphone library. However, in the article, the authors suggested that although there is no so-called best technique, the Metaphone seems like an efficient one for the English alphabet.

Thus, for our research, we will be using the Metaphone library to find the similarity between the given the domains.

III. Methodology

In this section, we explain the methodology to investigate whether children are the target of typosquatting and the research flow is referred to in the figure 3. We understand the type of errors children make of various ages as mentioned in the figure 3 for the sub-question 1, and we list the children domains mentioned as sub-question 2 in the figure 3. The listed domain is then fed to the tool DNSTwist to obtain the typo-variant domains from the original domain. Based on the obtained typo-variant domains, we are identifying the percentage of children errors covered in the DNSTwist, which is mentioned as sub-question 3 in the figure 3. Once the percentage of children errors are obtained, we are investigating whether there is any evidence of blacklisting concerning children's domain which is sub-question 4 in the figure 3; see the figure 3 for an overview of our research.

A. *Methodology for SQ1*

As explained in the literature review section II, a plethora of literature reading has been made to understand children's errors. As per the literature, all the writing, typing, and reading errors fall into addition, omission, and substitution categories. This methodology aims to identify the childrens errors. Some of the error categories are addition, omission based on phonetic similarity of words, and substitution of the intended characters with the confusing character. On the other hand, considering the QWERTY keyboard, the character adjacent to the intended character can also be added or replaced for the addition and substitution categories. In addition, the error categories are observed in the various age groups from 5 to 14 years. From the literatureII it has been observed that 5-10 years children make errors based on phonetic similarity of words with respect to addition, omission, and substitutions. On the other hand, 11-14 years children make error with respect to QWERTY keyboard.

However, the confusion character table is constructed for the substitution category, where the intended character is replaced with a confusing or visually similar character. The aim of developing the confusion character table is to record the characters that are confusing to children based on visually similar looking set of letters. For example, ’b’ and ’p’ are confusing letters for children ages 5-10 years. The following metrics are constructed to develop the table of confusion characters and are as follows:

1) **Numerical similarity**
   Children sometimes might gets confused with the letter to a number which looks visually similar while reading. Thus, while writing, and typing there is a possibility that they might write what they have interpreted while reading. So, there is a chance that children might replace the letter with the confusing number which looks similar visually. Examples of such cases are ’D’ and ’0’, ’B’ and ’8’, et cetera. Such confusing pairs will be recorded in the confusion character table.

2) **English letter similarity**
   Children sometimes might gets confused with the letter to another English letter that looks visually similar while reading. So, there is a chance that children might replace the letter with the confusing letter which looks similar visually. Example of such case is ’p’ and ’b’, ’n’ and ’u’, et cetera. Such confusing pairs will be recorded in the confusion character table.

3) **Special characters and other language similarities**
   For some of the set of letters, there could be other special characters that look visually similar, and one such example is ’S’ and ’$’. On the other hand, the few other language characters such as Spanish, et cetera, more or less look similar to certain English characters. An example of a such a case, the letters ”à”, ”á”, ”â”, ”ã”, and ”ä” look visually similar to

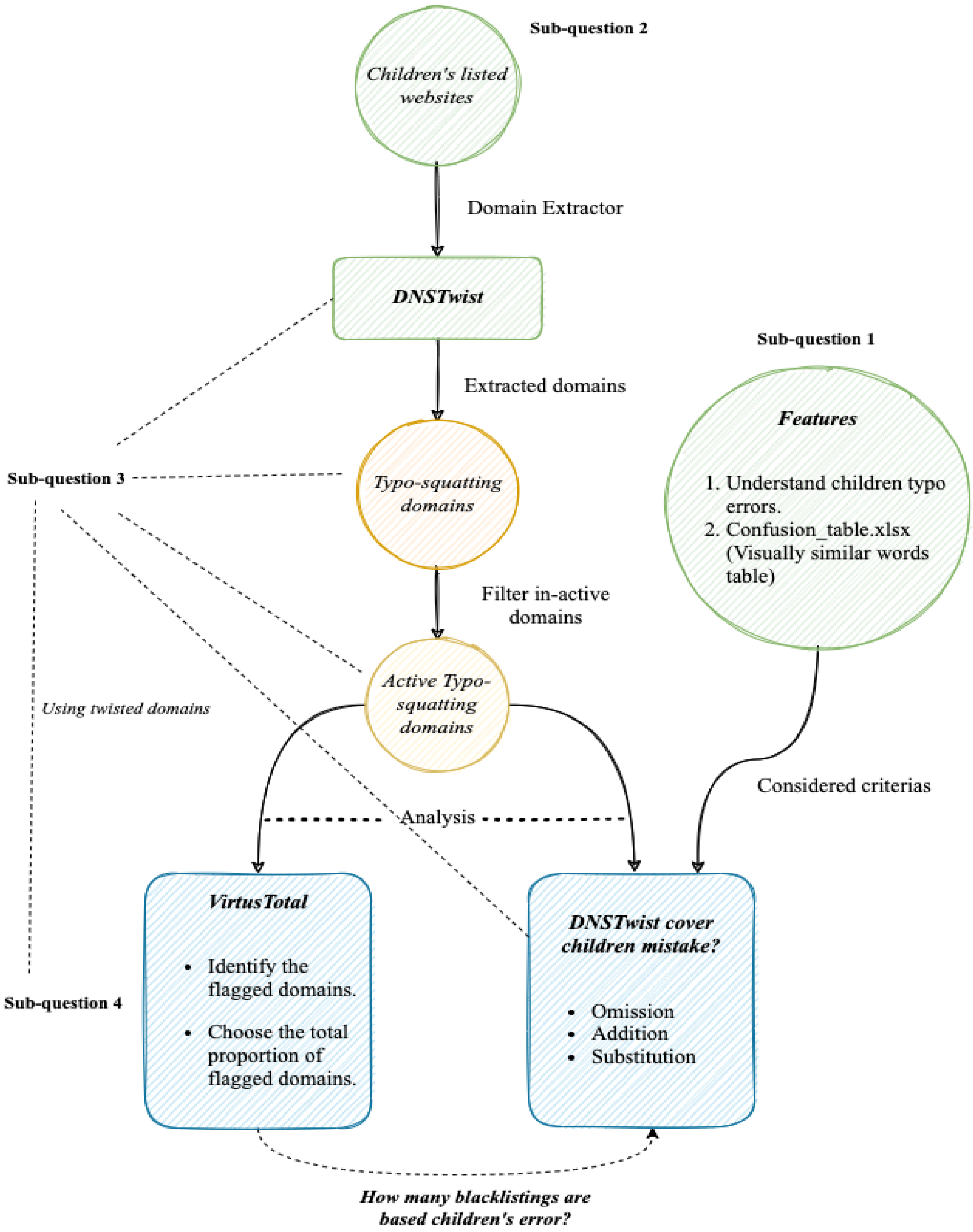

Sub-question 2
Children's listed websites
Domain Extractor
DNSTwist
Extracted domains
Sub-question 1
Features
1. Understand children typo errors.
2. Confusion_table.xlsx (Visually similar words table)
Sub-question 3
Typo-squatting domains
Filter in-active domains
Active Typo-squatting domains
Using twisted domains
Considered criterias
Analysis
VirtusTotal
Identify the flagged domains.
Choose the total proportion of flagged domains.
DNSTwist cover children mistake?
Omission
Addition
Substitution
Sub-question 4
How many blacklistings are based children's error?

Fig. 3. THE STUDY'S SETUP (DIAGRAM CREATED ONLINE [16])

the English letter 'a'. Such confusing pairs will be recorded in the table.

Thus, based on the mentioned methodology, the result would be to obtain the confusion character table for the children.

B. Methodology for SQ2

Once we obtain the categories of errors of children of various ages from the earlier SQ1 methodology III-A, we need to identify children's domains based on the popular contents for the further investigation. Thus in this section of **SQ2**, we are providing metrics to obtain the list of domains that would interest children.

The methodology decided to obtain domain list is mentioned below as points and it has three layers:

1) **Content-providers**
As stated in the article by Matrix [32], children do live in an on-demand world, where they expect to find their shows not only on TV but also everywhere else, such as online platforms like Netflix[2], Amazon prime, Now TV[3], where they can view even using mobiles, tablets, et cetera. Also, on online platforms such as Netflix, Youtube[4], Amazonprime[5], the content is distributed to a wide range of age groups, including children contents as well. Once the providers are chosen, they are assessed to identify the popular content for children of various ages. Some of the popular contents may possess an official domain that children might be interested in viewing as they may provide small videos, games, et cetera.

2) **Domains**
The popular contents are obtained from the earlier step. Once the contents are received, they will be assessed whether any official domains might contain engaging content for the children, such as games, videos, online shopping, et cetera.

3) **Suggestion from people**
Another option for this research is to check people's responses via public platforms. It may also be an important factor as they might know what domains their children often use. The public platforms might come in handy, and we are using Reddit[6]. Nevertheless, with consideration, Reddit can still be a terrific site to ask queries, seek help, or gain insight into people's lives all around the world. In the Reddit platform some of the communities were chosen to ask the questions about the children domains such as 'websiteserviceschildren', 'childrensbooks', 'teenagers', et cetera.

Thus, this outcome is that we will collect the list of domains of children's interest-based in several age groups.

C. Methodology for SQ3

Once we obtain the errors categories, the type of errors children make, and the list of domains of the children's interest, we will feed the domains on the sequence to the tool DNSTwist. Meanwhile, the working functionalities of the DNSTwist have been mentioned in the Literature review. Also, in the literature review, several possibilities of typosquatting are mentioned through the permutation technique of this tool.

The setup for this methodology is as per the figure 4 respectively. The following actions will taking to approach the problem:

- **Addition**
As discussed in the earlier steps III-A, the children in the 5-10 years category make errors mostly on phonetically similar words. Thus, as mentioned in the literature review, we will be using *Metaphone* library for this step. Initially, we will be taking all the Addition and Insertion category domains from the DNSTwist as compared domains. The original domain or the brand needs to be compared with the comparer domains, and as mentioned in the literature review, we will be using *Levenshtein distance* metric. If the domains are not similar, the Levenshtein distance will not be zero. If the value is non-zero, it can be passed to the Metaphone library to classify the similarity factor.
On the other hand, the children in the 10-14 years category most likely make errors using the keyboard layout. They will probably type the adjacent key of the intended letter. Thus, to find whether the key is adjacent to the intended letter or not, the metric *Euclidean distance* has been used. The algorithm is designed and developed considering the QWERTY keyboard layout. The result from the program would be that if the key is adjacent, it will result in the value 1.0, and if not, it will be a higher

[2] www.netflix.com
[3] www.nowtv.com
[4] www.youtube.com
[5] www.primevideo.com
[6] www.reddit.com

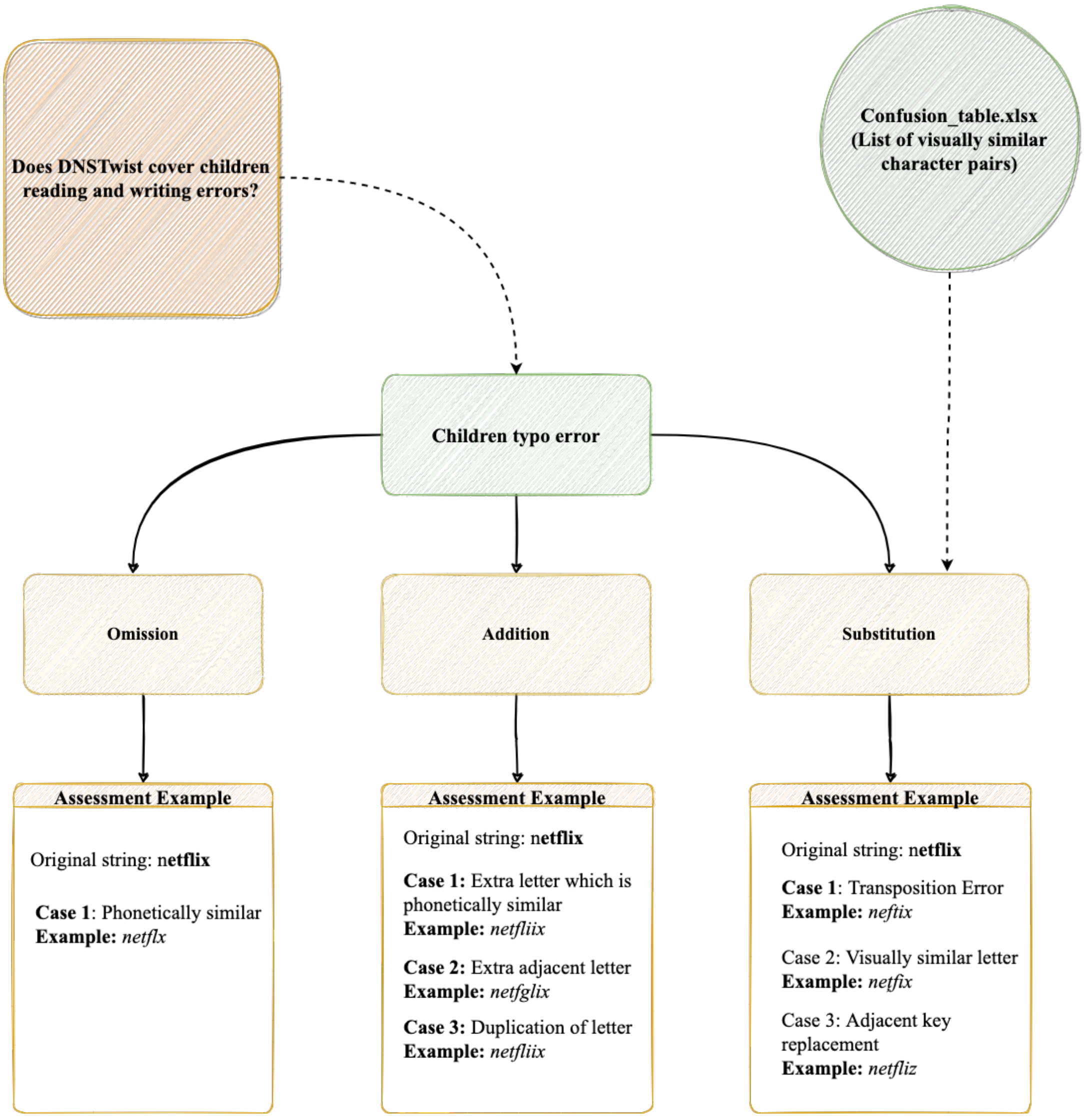

Does DNSTwist cover children reading and writing errors?
Confusion_table.xlsx (List of visually similar character pairs)
Children typo error
Omission
Addition
Substitution
Assessment Example
Original string: netflix
Case 1: Phonetically similar
Example: netflx
Assessment Example
Original string: netflix
Case 1: Extra letter which is phonetically similar
Example: netfliix
Case 2: Extra adjacent letter
Example: netfglix
Case 3: Duplication of letter
Example: netfliix
Assessment Example
Original string: netflix
Case 1: Transposition Error
Example: neftix
Case 2: Visually similar letter
Example: netfix
Case 3: Adjacent key replacement
Example: netfliz

Fig. 4. CHILDREN ERROR CATEGORIES TO BE EVALUATED IN THE DOMAIN LIST (DIAGRAM CREATED ONLINE [16])

value. The example of this case has been mentioned in the figure 4 of the Addition category respectively.

In a nutshell, the percentage of this error category can be obtained by the positive result to the overall comparison that has been made.

- **Omission**
  As discussed in the earlier Addition category, the children in the 5-10 years in this Omission category make errors most likely on phonetically similar words. As mentioned in the earlier category, yet again here as well, we will be using the *Metaphone* library to check the phonetic similarity. We will be taking out the Omission category domains from the DNSTwist as the compared domains for the investigation. The original domain or the brand needs to be compared with the comparer domains, and as mentioned in the literature review, we will be using *Levenshtein distance* metric. If the domains are not similar, the Levenshtein distance will be less than one for the omission. If the value is non-zero, it can be fed to the Metaphone library to classify the similarity factor.

  In a nutshell, the percentage of this error category can be obtained by the positive result to the overall comparison that has been made.

- **Substitution**
  In this part of the section, we will discuss the substitution category here, which seems more challenging than the other two mentioned. As mentioned in the **SQ1**, in the substitution category, the children in the 5-10 years category are more prone to make errors by replacing the intended character with a look-alike character or the confusing character. From the outcome of the method of the **SQ1**, we will have the table to investigate the substitution in the DNSTwist domains list. However, the **homoglyph** category is filtered from the DNSTwist to investigate. In the software, every character or letter can be expressed in it's **Unicode** as it is unique for all the characters. Based on the Unicode technique and the constructed confusion table, the algorithm is designed and developed to find the number of substitutions from the homoglyph category of DNSTwist.
  We have designed the algorithm for the **homoglyph** based comparison as it can be seen in algorithm blocks 9 and 2 respectively. An algorithm block is a set of instructions or lines written for completing a task or solving a problem. The main aim of this algorithm block is to identify the percentage substitution of confusion characters in the twisted domains from the DNSTwist based on the confusion characters table constructed from the sub-question III-B. For example, one of the confusing pairs for children is 'p' and 'q'. In the twisted domain for the original domain 'pokemon', we identify if 'p' is replaced by 'q' in any of the twisted domains and represented as 'qokemon'. However, this identification is performed by converting all the set of letters, characters to it's unicode representation which will help to identify accurately. In the algorithm block 9, it can be seen that all the characters from domains from twisted domain and characters from the confusion table are being represented in their Unicode.
  Also, in algorithm block 2, it can be seen that the Unicode obtained from twisted domains is compared against the Unicode of the confusion table. If the replacement in the domains is as per the confusion table, then we are increasing the counter by one.
  Overall, after the match is obtained and based on the number of comparison is performed, the percentage is calculated based on the following:

$$A = (number\ of\ substitutions\ obtained)$$

$$B = (number\ of\ comparisons\ performed)$$

$$Substitution\ percentage = \frac{A}{B}$$

  On the other hand, children of 10-14 years are prone to make errors concerning substituting the adjacent keys from the QWERTY layout instead of the intended character. Also, they could swap immediate letters in the word. To obtain the results of this category, we are using the *Damerau-Levenshtein* for assessing the transposition of characters and *Euclidean distance* to check whether the intended letter is being replaced by the adjacent key in the QWERTY layout.

  The outcome of this method is to obtain the percentage of substitution of the mentioned classified categories in the domains list from the DNSTwist for the mentioned children error categories.

D. *Methodology for SQ4*

From the outcome of the **SQ3**, we obtain the percentage of children error-based domains existing in the list obtained from the DNSTwist. However, as the figure 3 depicts, the percentage of blacklisted domains from the overall DNSTwist output will be recorded. To investigate whether the domains are blacklisted, we will be using the service from VirusTotal as mentioned in the literature review. VirusTotal is a good choice for the research as it provides the public APIs where the software programmers can access their end-point to obtain the result. The developer's guide from the VirusTotal has been used to integrate [7]. Another important factor and it is believed that VirusTotal provides an unbiased service to the users which has an aggregated results from several vendors

[7] https://developers.virustotal.com/reference/overview

Algorithm 1: ALGORITHM PART - I

```
Function Main:
    INITIALIZE: Original_domain = DNSTwist_original_domain
    INPUT: Confusion_table.xls
    /* Obtain Unicode value for each character from the confusion table */
    for each of the domain from DNSTwist tool do
        OBTAIN: Filter ”Homoglyph” domains
        for each domain from the Homoglyph category as ”comparer domains” do
            if Levenshtein distane(original domain, comparer domain) > 0 then
                /* The comparer domain is represented in their Unicode values */
                CALL: count = compare_unicode_of_string(original_domain,
                comparer_domain, confusion_table_data)
                /* Record count of replacement of each domain */
    /* Tabulate the replacement count obtained for all the domains */
    return 0
```

Algorithm 2: ALGORITHM PART - II

```
Def compare_unicode_of_string(original_domain, comparer_domain,
 confusion_table_data):
    INPUT: Original_domain, comparer_domain, confusion_table_data
    /* Initializing the parameters */
    SET: count = 0
    SET: count_comparison = 0
    SET: Col A = confusion_table_data[original_character], Col B =
    confusion_table_data[confusion_character]
    for each unicode in the original_domain do
        /* Comparing original domain character against col A of confusion table and
           comparer domain character against col B of confusion table */
        if unicode->original_domain ! = unicode->comparer_domain then
            count_comparison = count_comparison + 1
            SET: Indexes = ”unicode− >comparer_domain” in ”Col B” for each of the indices
            from earlier step do
                /* If the match is found as per the confusion table with respect to
                   original and comparer domains, the counter is incremented */
                if unicode->original_domain == Col A and unicode->comparer_domain ==
                 Col B then
                    count = count + 1
    /* Returning the values to the Algorithm Part - I */
    return count_comparison, count
```

such as Spam Haus[29], URL Haus [27], et cetera. The blacklist percentage of each brand or original domain is calculated using the following formula:

$$A = (number\ of\ blacklisted\ domains)$$

$$B = (Overall\ domains\ from\ the\ DNSTwist)$$

$$Blacklist\ percentage = \frac{A}{B}$$

However, not all the blacklisted domains are based on the children error category. Thus, the blacklist percentage for each of the error categories mentioned in the earlier section **SQ1** is determined. On the other hand, the vendor's percentage for each original brand that voted the domain as malicious is recorded. Also, we will record the threat categories of each of the blacklisted domains.

E. Methodology for the RQ

The final methodology for **RQ** is the aggregated methodology of all the **SQ** sub-questions. Alongside, it is important to have a comparison against top records such as Alexa[33]. The comparison can depict whether the children are the target of typosquatting as typosquatting mimics the children's errors.

IV. RESULTS

This section presents the results of each of the research questions adopting the methodologies mentioned in the earlier sections III.

*A. **Results for SQ1***

As mentioned in the methodology III-A, the literature reading has been done to understand children's several errors at various ages. The literature reading also noted the error categories such as Addition, Omission, and Substitution. The age groups that are prone to make errors in various categories are tabulated in the table IV respectively.

TABLE IV
ERROR CATEGORIES OF VARIOUS AGES

Age group	Errors
5-10 years	• **Addition, and Omission:** Phonetically similar errors • **Substitutions:** Confusion character/Visually similar character replacement.
11-14 years	• **Addition:** Adjacent key addtion • **Substitutions:** Transposition of characters, Adjacent key replacement

To summarize table IV, the children of age groups 5-10 are prone to make mistakes on the words that sound similar to another. Thus, based on that trait, the letter or character from the word might be added or even removed. For the same age group, they are prone to replace the character similar to another character in terms of a look-alike. The table is constructed for the investigation and is tabulated and mentioned in the Appendix A. The example of the confusion character table are 'b' and 'p', 'b' and 'd', 'O' and '0', 'S' and '$', etc.

On the other hand, teenagers aged 11-14 are prone to make mistakes based on the typing QWERTY keyboard layout. Where while typing, the adjacent key to the intended character might be added or replaced.

*B. **Results for SQ2***

Once the error categories of children of various age groups are obtained from the earlier step IV-A, the next step would be to obtain the list of domains of children's interest. Meanwhile, the methodology to obtain the list of domains is mentioned and discussed in the section Methodology and sub-section for SQ2 III-B.

As the analysis conducted by the Ampere [34], Netflix has 40% of the contents about the 0 to 6 years children group in the year 2017, and it dipped to 35% in the year 2019. There are 34% of the contents in both 2017 and 2019 for 7 to 9 years. Also, there are 26% of contents in 2017 and 30% of contents in 2019 for 10 to 18 years old.
On the other hand, with the popular content provider, Amazon prime, there are 56% of the content about the 0 to 6 years children group in 2017, and it dipped to 52% in 2019. There are 23% of the contents in 2017 and 21% of the contents in 2019 for 7 to 9 years. Also, there are 26% of contents in 2017 and 22% of contents in 2019 for 10 to 18 years old. However, with the popular content provider Now TV, there are 35% of the content about the 0 to 6 years children group in 2017, increasing to 39% in 2019. There are 23% of the contents in 2017 and 36% of the contents in 2019 for 7 to 9 years. Also, there are 30% of contents in 2017 and 25% of contents in 2019 for 10 to 18 years old. Thus, considering the above statistics it is important to investigate for the popular contents from such providers.

These days children are more likely to use the content-providers such as Netflix, Primevideo, et cetera during their leisure. They must have drawn the attention to popular content of their interest, and if they see such a name in an advertising link, there is a high possibility that they could click such a link.

Thus, popular content providers such as Netflix, Primevideo, et cetera are assessed to identify the popular content for several age groups. Based on the popular content of several age groups, the investigation is performed to check whether the content possesses any official domain. Based on the popular contents like 'angrybirds','mrbean', 'pokemon', et cetera, we assessed if they have any original domains by their names, and such domains are noted in the list. With this investigation, we listed 20 pretty popular domains. However, on a minimal number, the public response has been recorded as well from the platform Reddit[8]. As mentioned in the methodology 3, the communities such as 'websiteserviceschildren', 'childrensbooks', 'teenagers' were asked about the domains that children use regularly, and we received the list of 5 domains in response from a couple of people. In overall, the list of 25 domains are considered for the experiment and are mentioned in the Appendix B.

C. *Results for SQ3*

Once the error categories of children and the list of domains that children use are obtained from IV-A, and IV-B, it is important to provide the list of domains to the DNSTwist, the typosquatting tool, to investigate the amount of children's error types being covered by the tool. Based on the mentioned methodology for the research sub-question 3, III-C, the algorithms are developed to identify the overall error percentages covered by DNSTwist based on the categories Addition, Omission, and Substitution for the target children group of 5-10 years and 11-14 years.

Thus, based on the libraries chosen, the algorithm is developed to quantify the results which depicts the overall percentage of errors of children's category covered by the DNSTwist. The percentages of several categories can be seen in the tabulated table V.

To summarize the table, based on the error categories are as follows:

TABLE V
PERCENTAGE OF CHILDREN ERROR OF VARIOUS CATEGORIES COVERED BY DNSTWIST

Error types	Description	Percentage
Substitution error (Visually similar/confusing character error)	This error occurs due to replacing the intended character with another character that is visually similar to each other. The similarity list or the confusion table is provided for the assessment. Example: ‘primevideo’ and ‘drimevideo’.	4.84%
Omission error	This error occurs due to removing the character and phonetically sound similar after the removal.	82.89%
Addition error (Phonetic similarity)	This error occurs due to the addition/insertion of the character and phonetically sounding similar after the addition.	78%
Substitution error (Adjacent keys)	This type of error occurs when an intended key is replaced by the adjacent keys in the keyboard layout. Example: 'Netflix' and 'neylix'.	89.1%
Transposition error	The transposition error occurs when the immediate characters are swapped in the given the word or the website.	88.9%
Addition error(addition of adjacent key)	This error occurs due to the insertion of additional characters adjacent to the intended character in the keyboard layout. Example: 'Netflix' and 'Netfliox'	55%

- **Addition**
  The children of 5-10 years of age group are prone to make errors based on phonetic similarity. We designed and developed an algorithm using the libraries *Levenshtein* distance and *Metaphone* libraries. However, based on the metrics decided, we obtained the addition error of 78%.

  On the hand, teenagers in the 11-14 age group are prone to make errors such as adding or inserting adjacent keys to the intended character based on the QWERTY keyboard layout. Thus, as mentioned in the methodology, the metrics used are *Euclidean distance* and the result obtained is 55%.
- **Omission**
  The children of 5-10 years age group are prone

[8] www.reddit.com

to make errors that are likely on phonetic similarity. To investigate we used the libraries *Levenshtein distance* and the *Metaphone*. However, based on the chosen metrics, we obtained the result of 82.89%.

- **Substitution**
  The children in the 5-10 years categories are more prone to errors by replacing the intended character with a visual look-alike or confusing character. Example 'b' and 'p'. Thus, on comparing the number of replacements that are similar to the confusion table constructed against the **homoglyph** category of DNSTwist, we obtained the overall percentage of 4.84%, which is the least compared to other categories. We have designed the algorithm for the **homoglyph** based comparison as it can be seen in algorithm blocks 9 and 2 respectively. In the algorithm block 9, it can be seen that all the characters from domains from DNSTwist and the confusion table characters are represented in their Unicode.
  Also, in the algorithm block 2, it can be seen that the Unicode representation of domains from DNSTwist is compared against the Unicode representation of the confusion table. If the replacement in the domains is as per the confusion table, then we are incrementing the counter by one.
  Overall, after the match is obtained and based on the number of comparisons is performed, the percentage is calculated based on the following:

$$A = (number\ of\ substitutions\ obtained)$$

$$B = (number\ of\ comparisons\ performed)$$

$$Substitution\ percentage = \frac{A}{B}$$

As observed in the figure 5, the replacement count based on the confusion table constructed for all the chosen domains are listed. It can be inferred from the figure that content-provider domains have relatively more replacement counts compared to the other domains. Also, based on the mentioned formula above, the percentage of substitution for each domain is recorded and it can be seen 6.
Another interesting observation is that even though there is a replacement count of 14 for the domain *play.barbie.com* from the figure 5, the overall percentage is still 0 from the figure 6 as the number of comparison performed for each of the typo variant domain from the DNSTwist is relatively more. Similarly, on contrary, even though the substitution count for the domain *marvel.com* is 3, the overall percentage is average in the table. However, the percentage of substitution for the all the domains is relatively lower and recorded to 4.84% in the table V.

However, looking at the error categories and the error pattern, it is more of a normal typosquatting as children's age group increases.

D. *Results for SQ4*

From the earlier result sections IV-A,IV-B, and IV-C, we have obtained the children's error categories, the list of domains of children's interest, and the percentage of error categories covered by the DNSTwist. It is important to identify how many of the twisted or permuted sites obtained by the DNSTwist are blacklisted. As mentioned in methodology, we are using the service **VirusTotal** [28] to identify the evidence of blacklisting. VirusTotal is a reliable and trustable service as it is an aggregation of several vendors such as Nortan, McAffee, Spam Haus, et cetera.
Based on the VirusTotal API, we developed an algorithm to identify the blacklisted domains. The overall number of blacklisting domains obtained is tabulated in the table VI.

TABLE VI
COUNT AND PERCENTAGE OF BLACKLISTING

Measurement	Domains	Blacklisted domains	Percentage
10 days	2363	248	10.49%
20 days	2363	249	10.53%
30 days	2363	253	10.57%

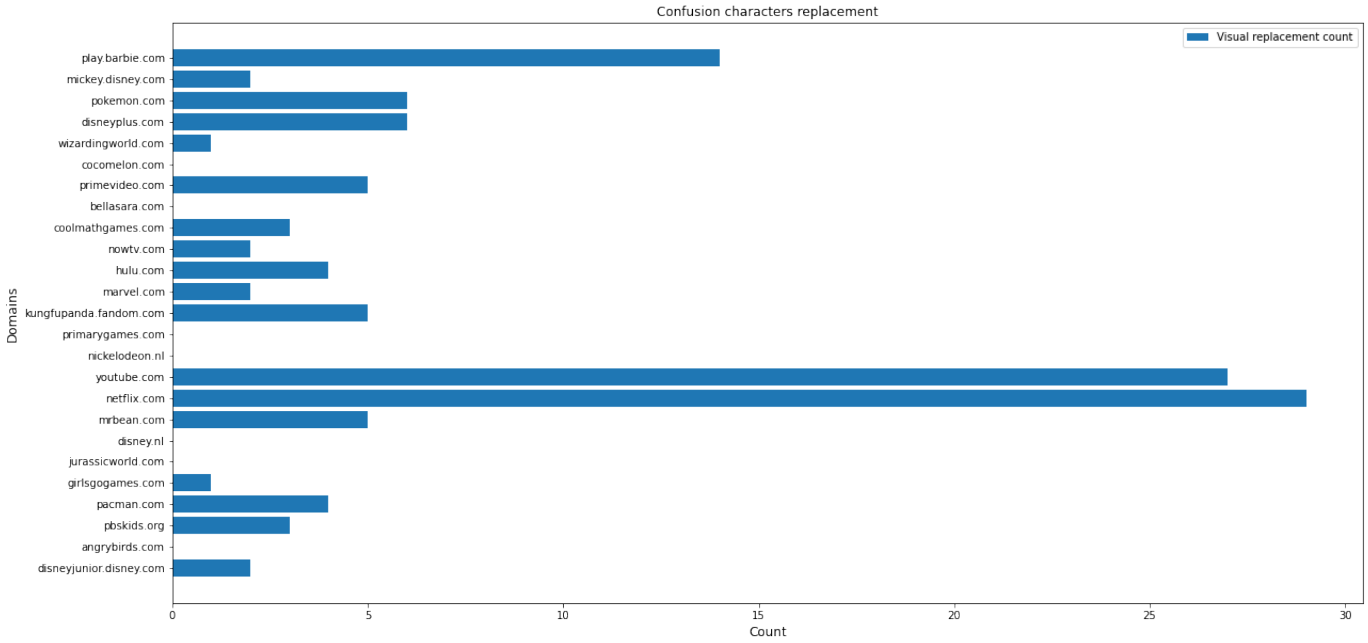


Fig. 5. SUBSTITUTION COUNT BASED CONFUSION CHARACTER TABLE (BAR GRAPH CREATED USING MATPLOTLIB [35])

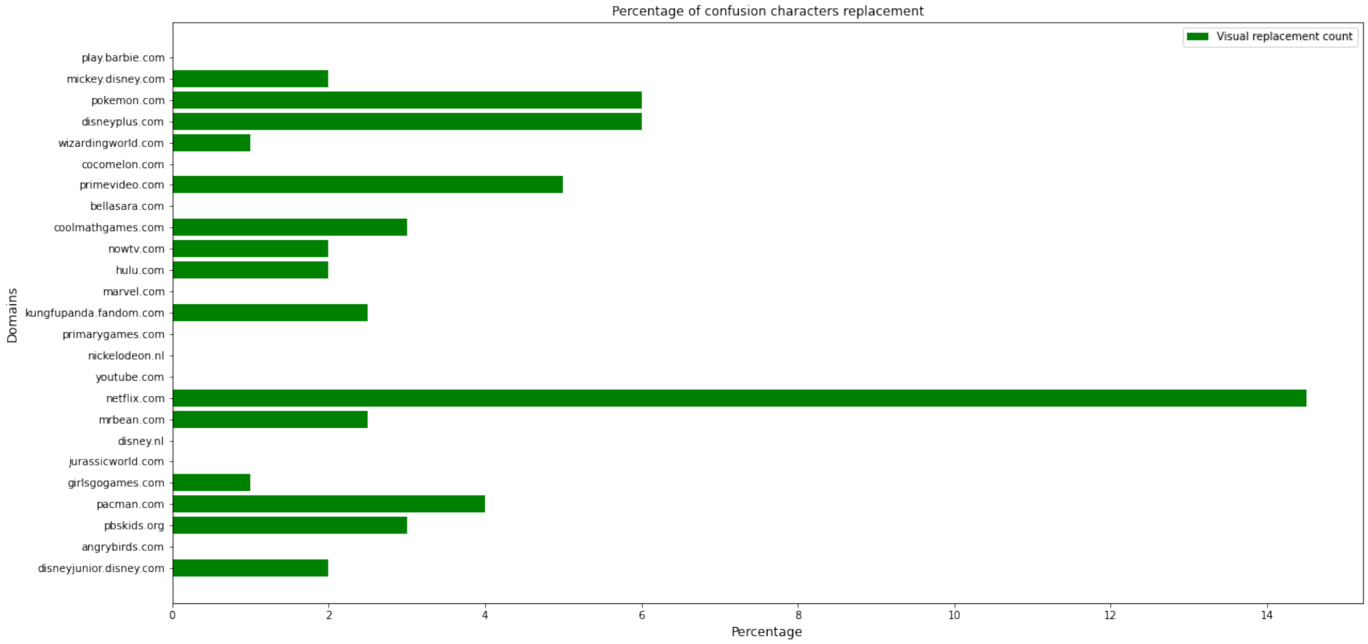


Fig. 6. SUBSTITUTION PERCENTAGE BASED CONFUSION CHARACTER TABLE (Bar graph created using matplotlib [35])

As it can be seen in the table referred VI, it can be inferred that overall, 2363 domains were obtained from the DNSTwist considering the domain list that was constructed for the experiment. Among them, 248 domains are blacklisted as malicious, which constitutes 10.49%. On the other hand, within the next 10 days, the number of blacklisting increased by one leading to 10.53%. Also, with the subsequent 10 days, the blacklisting number was increased by four, leading to 10.57%. Thus, it can be depicted that the continuous measurement is appearing to be stable. However, the overall distribution count of the malicious domains can be seen in the figure 7, where the content providers such as Netflix and youtube are the top two in the list and other children domains are at the bottom of the list. The overall blacklisting number changes for each of the domains that are used for this research are represented in the figure 8.

The count of blacklisting for each domain based on DNSTwist permuted domains and it's percentage can be seen in the figures 9, and 10 respectively.

However, all the domains listed in the table VI are based on the typo variant domains listed by the DNSTwist. Thus the filter category of ***Addition, Omission, and Substitution*** categories are applied to identify the blacklisting count of each of the children categories mentioned in the results IV-C and the fragmented count for each category is recorded in the table VIII. Thus, out of 253 blacklisted domains, 106 blacklisted errors are based on children error categories which is referred in the table VII. Also, even though the measurement was carried out for about 30 days, the blacklisting count for any children's error categories remained unchanged.

TABLE VII
COUNT AND PERCENTAGE OF BLACKLISTING BASED CHILDREN ERROR

Domains	Blacklisted domains	Percentage
1006	106	10.536%

TABLE VIII
BLACKLISTED DOMAIN BASED ON CHILDREN ERROR CATEGORIES

Error types	Domains	Blacklisted domains	Percentage
Substitution error (Visually similar/confusing character error)	186	32	17.204%
Omission error	215	23	10.69%
Substitution error (Adjacent keys)	316	26	8.22%
Substitution error(Transposition error)	102	10	9.80%
Addition error(addition of adjacent key)	187	15	8.02%
All error categories	**1006**	**106**	**10.536%**

The overall count of the vendors from VirusTotal that raised the domains as malicious are also noted and it can be seen in the referenced figure 11 respectively. Thus, it can be seen that the most blacklisting were found with the content providers rather than with the other domains.

Another important factor to consider is that sometimes the original brand, such as Netflix, youtube registers the typo variant domains by themselves as a defensive mechanism against the cybercriminals, which is typically known as ***protective domain registration***. Also, the malicious actors or typosquatters will be unable to register these domains in the future because of protective domain registration. From the mentioned table VI, we have determined the overall number of blacklisting based on the typo variant domains from the DNSTwist. To investigate whether any of the domains have been registered by the original brand, the ***WHOIS*** [38] is used. The name WHOIS is a short form of ”Who is responsible for this domain name?”, which has the details such as registered users, created date, modified date, registry, registrant, registrar, et cetera. The property considered here to compare is the ”registrar” from WHOIS. The registrant is the domain's legal owner who registered it. The registry is the organization in charge of maintaining a list of domain names. Between the registry and the registrant, the registrar serves as a go-between. We compared the typo variant domains against the original brand and depicted the possible defensive domain registrations based on the registrar information. On performing WHOIS query for all the listed domains from the result section IV-B, we identified possible defensive domain registrations for several domains such as Netflix, Primevideo, Marvel, et cetera, which are all listed in the appendix C.

The count of defensive domain registration and blacklisting count is listed in table IX.

TABLE IX
DEFENSIVE REGISTRATION DOMAINS AND BLACKLISTING COUNT

Overall domains	overall blacklistings	Defensive domain registrations	Blacklisting defensive domains	percentage
2363	253	205	7	2.766%

From table IX, it can be seen that the overall domains are 2363 and out of which 253 were blacklisted, constituting 10.57%. From the WHOIS query, it is determined that 205 possible domains are protective domain registrations against the cybercriminals by the original brands and out of which

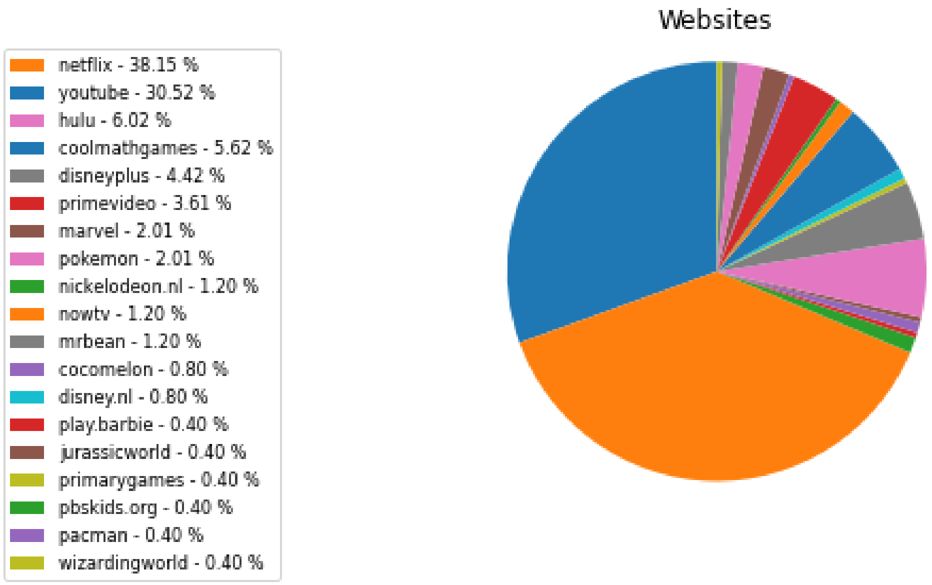


Fig. 7. OVERALL MALICIOUS DOMAINS DISTRIBUTION (PIE CHART CREATED USING MATPLOTLIB [36])

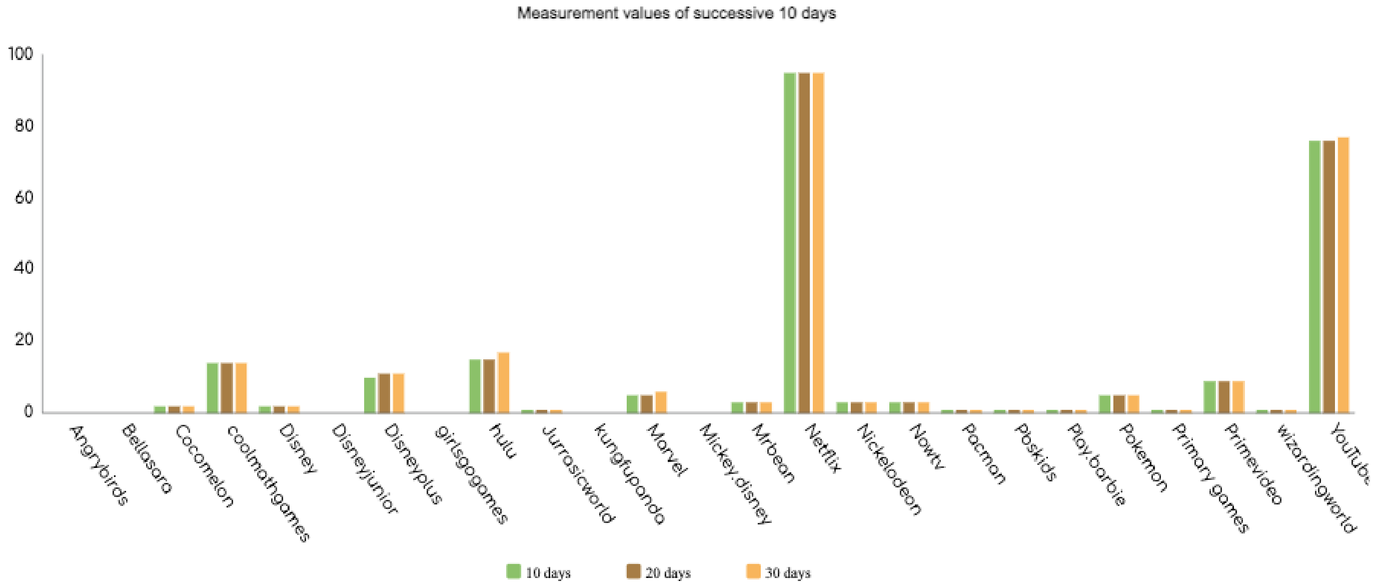


Fig. 8. MEASUREMENT OF DOMAINS FOR 30 DAYS (BAR GRAPH CREATED ONLINE [37])

7 are blacklisted as malicious, which results in 2.766%.

E. *Results for RQ*

The earlier mentioned measurement of evidence of blacklisting from VirusTotal has been carried out consistently for over 30 days. However, a slight variation in terms of overall percentage was recorded, and more or less, it appeared to be constant.

Alongside, the final comparison has been made against the Alexa top records [33]. The logic behind selecting such domains is that typosquatters will naturally target the most popular domain names to enhance their chances of attracting unwary visitors. The Alexa top 25 records showed more twisted or permuted sites from DNSTwist

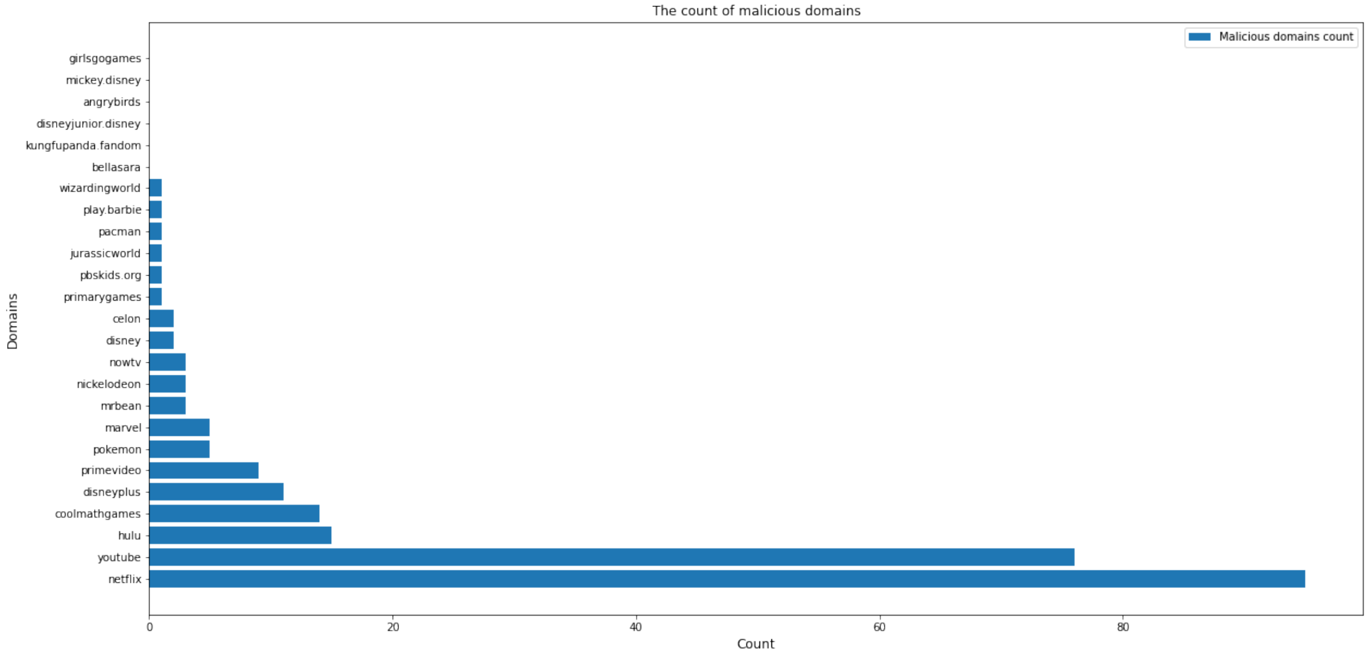


Fig. 9. BLACKLISTING COUNT FOR EACH OF THE DOMAIN BASED ON TWISTED SITES FROM DNSTWIST (BAR GRAPH USING MATPLOTLIB [35])

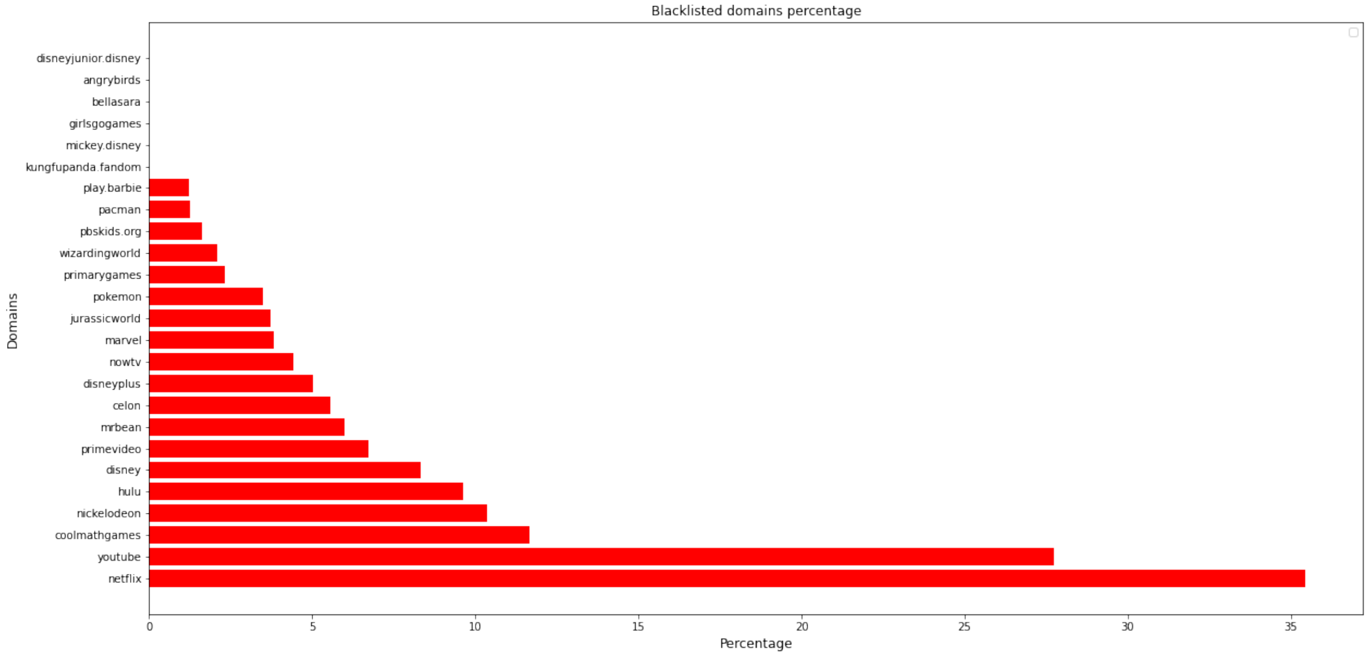


Fig. 10. BLACKLISTING PERCENTAGE FOR EACH OF THE DOMAIN BASED ON TWISTED SITES FROM DNSTWIST (BAR GRAPH USING MATPLOTLIB [35])

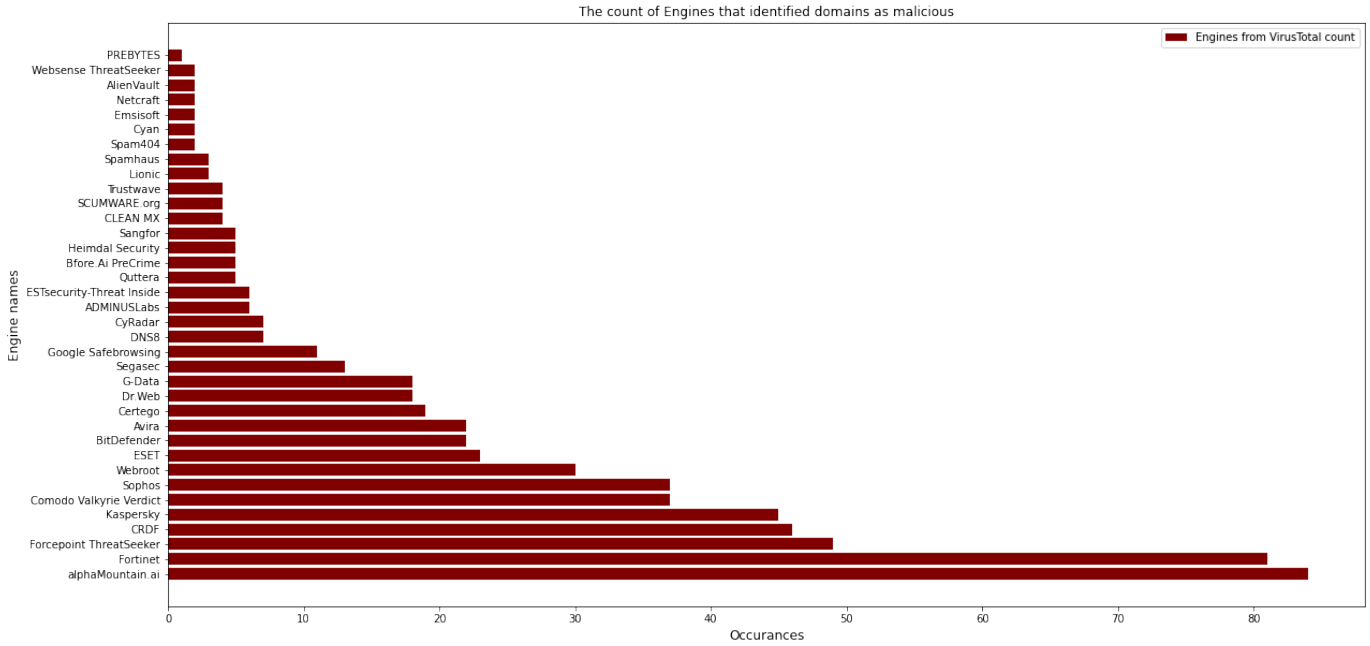


Fig. 11. VENDORS COUNT THAT RAISED DOMAINS AS MALICIOUS (BAR GRAPH USING MATPLOTLIB [35])

than the list of children's domains listed for this research. However, the number of blacklisting for the same 25 records is also relatively higher than the blacklisting from the children's domains.

Then Domain details of Alexa records are tabulated in the table X.

TABLE X
ALEXA TOP RECORDS DETAILS

Category	Domains	Blacklisted domains	Percentage
Alexa top 25	4105	822	20.02%
Alexa top 50	7193	1399	19.44%
Alexa top 100	12488	1968	14.95%

V. CONCLUSION

This research aims to find whether children are the target of typosquatting by understanding the typosquatting domain generation techniques, attacks, et cetera. In addition, when it comes to children, we understood the features such as their typical reading and writing errors which leads to typing errors, and derived the most prolific error categories. We investigated if the typosquatting tools cover the children's errors and, if it covers, what overall percentage is covered. Also, we investigated the evidence of blacklisting of typo variant domains obtained from the typosquatting tool with and without filtering children error categories.
One of the conclusions is that looking at the error patterns from the children of specific age groups, it seems like as the children's age group increases, it mimics the normal typosquatting. On the other hand, by comparing the domains with the Alexa [33] records and measuring for about 30 days, it depicts that children are significantly less threat to the typosquatting compared to the top records Alexa[33] as the percentage of blacklisting with Alexa[33] records is relatively way higher. The measurement for 30 days yielded consistent results in blacklistings for the listed domains for this research. Another evidence is based on the percentage of substitution obtained for the category 5-10 years, which was close to 17.204%, which also depicts that the potential homograph attack on the children is substantially low.

VI. LIMITATIONS AND FURTHER DISCUSSIONS

- The confusion character table considered is predominantly English alphabets, numbers, special characters, and a small number of alphabets from languages such as Czech and Spanish. However, the list can be enhanced further based on the different languages and phonetics.

- Also, it is evident from the literature that Dyslexia people also make the same mistakes with the words that are phonetically similar to other words. Thus, it can be evaluated to identify the distinct difference in detail.
- Also, concerning forming the domains list for research question 2, we did not consider the children's interests based on the geographic location and the language they speak. We constructed the table in general based on the popular contents. However, slight variant versions of domains such as more children's gaming sites can be included.
- Moreover, we did not dive into the content of the malicious domains, which could be anything, for example, advertisements, et cetera. So, it is important to have parental control to evaluate what their children are clicking and watching.
- Another important factor relating to content is obtaining the ethical board's permission, evaluating the content present in the blacklisting domains, and differentiating or categorizing them accordingly.
- Also, it is important to have a service in a device where the blacklisted domains list is maintained in the modern browsers to warn when the children are about to land on one malicious domain's page.

References

[1] *www.understood.org. (n.d.). Learning challenges that can impact typing*, vol. 13, p. 2021, Dec. [Online]. Available: https://www.understood.org/articles/en/learning-challenges-that-can-impact-typing

[2] W. T. Siok and C. Y. Liu, "Differential impacts of different keyboard inputting methods on reading and writing skills," *Scientific Reports*, vol. 8, 11 2018.

[3] J. Spaulding, S. Upadhyaya, and D. Mohaisen. The Landscape of Domain Name Typosquatting: Techniques and Countermeasures, 2016.

[4] J. Spaulding, D. Nyang, and A. Mohaisen, "Understanding the effectiveness of typosquatting techniques," *Proceedings of the fifth ACM/IEEE Workshop on Hot Topics in Web Systems and Technologies - HotWeb '17*, 2017. [Online]. Available: http://seal.cs.ucf.edu/doc/17-hotweb-ts.pdf

[5] H. E. Jochmann-Mannak, T. Huibers, L. R. Lentz, and T. Sanders. Children were searching information on the Internet: Performance on children's interfaces compared to Google., 2010.

[6] L. S. Mark, D. Shankweiler, I. Y. Liberman, and C. A. Fowler, "Phonetic recoding and reading difficulty in beginning readers," *Memory Cognition*, vol. 5, pp. 623–629, 11 1977.

[7] M. E. Swearingen, "When children make mistakes in spelling," *Elementary English, [online]*, vol. 29, no. 5, pp. 258–262, Dec. 1952. [Online]. Available: https://www.jstor.org/stable/41383950

[8] L. Hughes and A. Wilkins, *Typography in children's reading schemes may be suboptimal: Evidence from measures of reading rate*, vol. 23, no. 314-324, pp. 1467–9817, 2000.

[9] S. Walker and L. Reynolds, "Serifs, sans serifs and infant characters in children's reading books," *Information Design Journal*, vol. 11, pp. 106–122, 01 2003.

[10] A. Wilkins, R. Cleave, N. Grayson, and L. Wilson, "Typography for children may be inappropriately designed," *Journal of Research in Reading*, vol. 32, pp. 402–412, 11 2009.

[11] I. C. Simpson, P. Mousikou, J. M. Montoya, and S. Defior, "A letter visual-similarity matrix for latin-based alphabets," *Behavior Research Methods*, vol. 45, pp. 431–439, 10 2012.

[12] J. Read, S. MacFarlane, and C. Casey, "Measuring the usability of text input methods for children," citeseerx.ist.psu.edu, 2001. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.597.4830&rep=rep1&type=pdf

[13] J. Read and M. Norton, "Perspectives on hci research with teenagers — springerlink," link.springer.com, 2006. [Online]. Available: https://link.springer.com/content/pdf/10.1007%2F978-3-319-33450-9.pdf

[14] W. White, *Typing for Accuracy*. H. M. Rowe Company, Balitimore, 1932.

[15] Y.-M. Wang, D. Beck, J. Wang, C. Verbowski, B. Daniels, and M. n. Research. Strider Typo-Patrol: Discovery and Analysis of Systematic Typo-Squatting. [online] Available at. [Online]. Available: https://www.usenix.org/legacy/event/sruti06/tech/full_papers/wang/wang.pdf

[16] Draw.io, "Flowchart maker online diagram software," app.diagrams.net, 2021. [Online]. Available: https://app.diagrams.net/

[17] A. Banerjee, D. Barman, M. Faloutsos, and L. N. Bhuyan, *Cyber-Fraudis One Typo Away*. In IEEE INFOCOM, 2008.

[18] R. Yazdani, O. Van Der Toorn, and A. Sperotto, "A case of identity: Detection of suspicious idn homoglyph domains using active dns measurements." [Online]. Available: https://www.tide-project.nl/papers/eurospw2020.pdf

[19] M. Wolff and S. Wolff, "Attacking neural text detectors." [Online]. Available: https://arxiv.org/pdf/2002.11768.pdf

[20] N. Nikiforakis, S. Van Acker, W. Meert, L. Desmet, F. Piessens, and W. Joosen, "Bitsquatting," *Proceedings of the 22nd international conference on World Wide Web - WWW '13*, 2013.

[21] M. Ulikowski, "elceef/dnstwist," GitHub, 09 2020. [Online]. Available: https://github.com/elceef/dnstwist

[22] Y. Wang, D. Beck, and J. Wang, *Strider typo-patrol: discovery and analysis of systematic typo-squatting*. USENIX SRUTI, 2006.

[23] "www.typodomains.com. (n.d.)," *Domain Investigation Monitoring — TypoDomains.com. [online] Available at: [Accessed*, vol. 13, p. 2021, Dec. [Online]. Available: https://www.typodomains.com/

[24] T. Phuong Thao, A. Yamada, and A. Kubota, "Empirical analysis of domain blacklists," 01 2020.

[25] J. Van Der Velden, "Blacklist, do you copy? characterizing information flow in public domain blacklists." [Online]. Available: https://essay.utwente.nl/80567/1/Velden_BA_EEMCS.pdf

[26] V. , "Vx vault," vxvault.net. [Online]. Available: http://vxvault.net/ViriList.php

[27] A. Ch, "Urlhaus — malware url exchange," urlhaus.abuse.ch. [Online]. Available: https://urlhaus.abuse.ch/

[28] V. Total, "Virustotal," Virustotal.com, 2000. [Online]. Available: https://www.virustotal.com/

[29] S. Haus, “The spamhaus project,” www.spamhaus.org. [Online]. Available: https://www.spamhaus.org/

[30] E. Moreau, F. Yvon, and O. Cappé, “Robust similarity measures for named entities matching,” pp. 593–600, 2008. [Online]. Available: https://aclanthology.org/C08-1075.pdf

[31] K. Koneru, V. S. V. Pulla, and C. Varol, “Performance evaluation of phonetic matching algorithms on english words and street names - comparison and correlation,” *Proceedings of the 5th International Conference on Data Management Technologies and Applications*, 2016. [Online]. Available: https://www.scitepress.org/papers/2016/59263/59263.pdf

[32] S. Matrix, “The netflix effect: Teens, binge watching, and on-demand digital media trends,” *Jeunesse: Young People, Texts, Cultures*, vol. 6, pp. 119–138, 2014. [Online]. Available: https://www.researchgate.net/publication/270665559_The_Netflix_Effect_Teens_Binge_Watching_and_On-Demand_Digital_Media_Trends

[33] A. Top Sites, “Alexa top 500 global sites,” Alexa.com, 2020. [Online]. Available: https://www.alexa.com/topsites

[34] A. Analysis, “Kids age ranges - netflix, amazon, now tv - infogram,” www.ampereanalysis.com. [Online]. Available: https://infogram.com/kids-age-ranges-netflix-amazon-now-tv-1hnp27j1go8y2gq

[35] M. Python Library, “matplotlib.pyplot.bar — matplotlib 3.5.0 documentation,” matplotlib.org. [Online]. Available: https://matplotlib.org/3.5.0/api/_as_gen/matplotlib.pyplot.bar.html

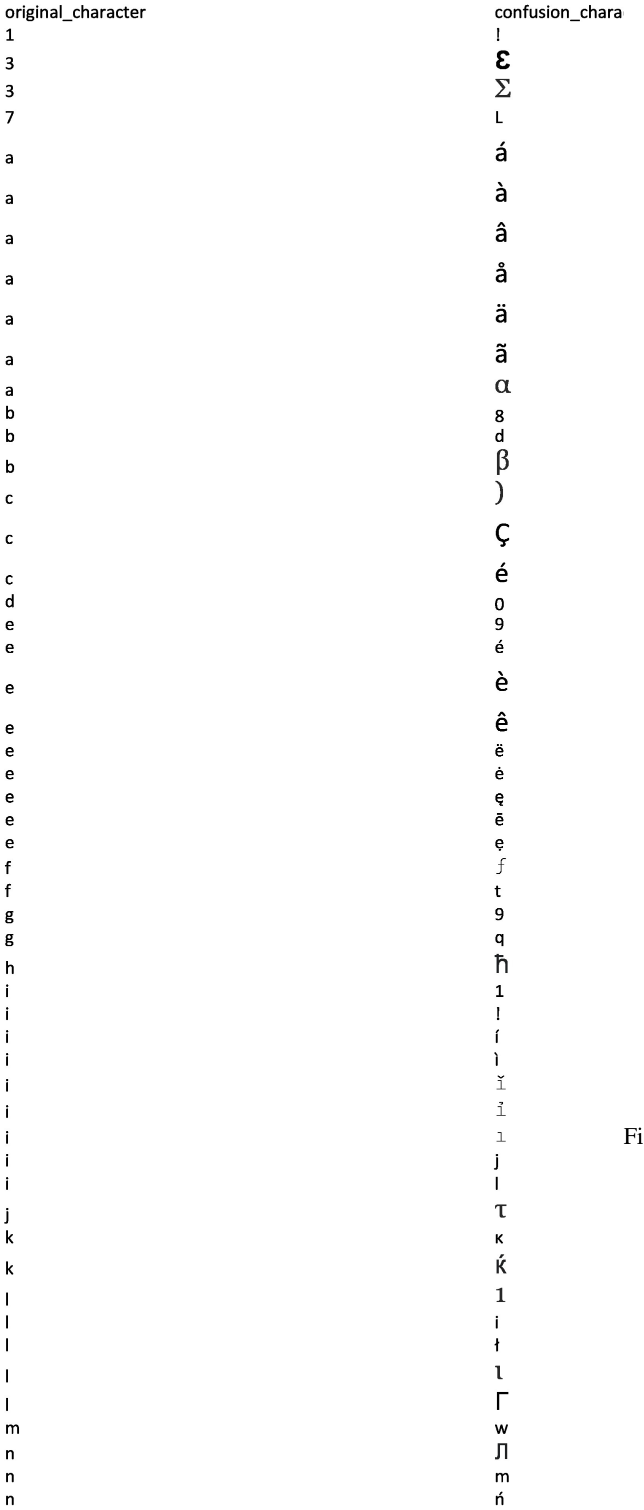
[36] P. Chart, “Basic pie chart — matplotlib 3.3.4 documentation,” matplotlib.org. [Online]. Available: https://matplotlib.org/stable/gallery/pie_and_polar_charts/pie_features.html

[37] V. , “Visme,” dashboard.visme.co. [Online]. Available: https://dashboard.visme.co/

[38] W. Wikipedia, “Whois,” Wikipedia, 11 2020. [Online]. Available: https://en.wikipedia.org/wiki/WHOIS

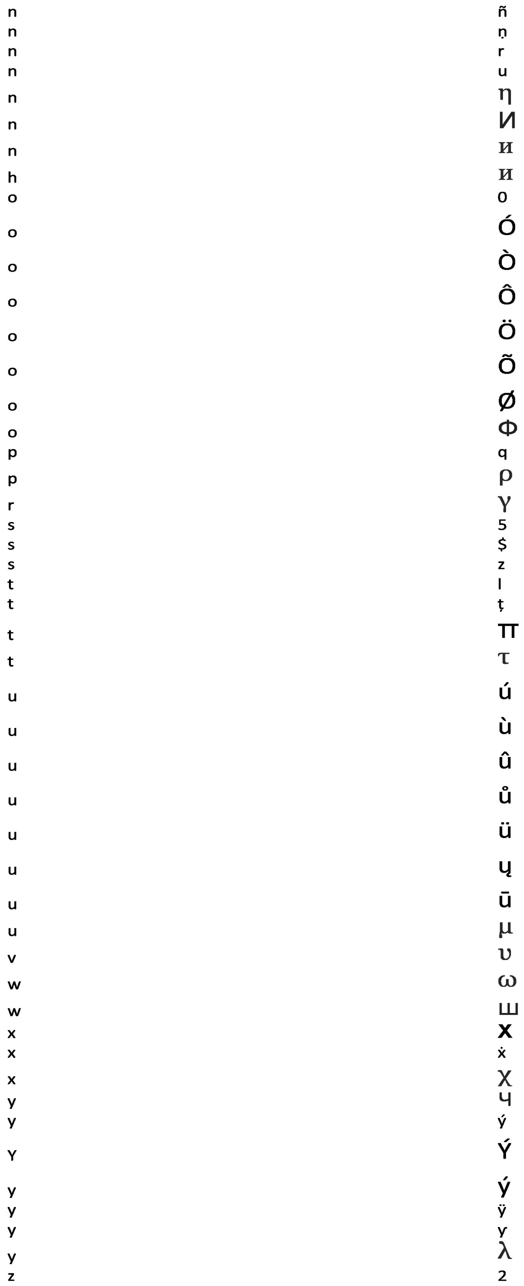
Appendix A
Confusion Table data

Confusion table data constructed for the research The figures 12, and 13 respectively.



original_character	confusion_chara
1	!
3	ε
3	Σ
7	L
a	á
a	à
a	â
a	å
a	ä
a	ã
a	α
b	8
b	d
b	β
c	)
c	Ç
c	é
d	0
e	9
e	é
e	è
e	ê
e	ë
e	ė
e	ę
e	ē
e	ẹ
f	ƒ
f	t
g	9
g	q
h	ħ
i	1
i	!
i	í
i	ì
i	ǐ
i	ỉ
i	ı
i	j
i	l
j	τ
k	κ
k	Ḱ
l	1
l	i
l	ł
l	ι
l	Γ
m	w
n	Л
n	m
n	ń

Fig. 12. Confusion table data part - I

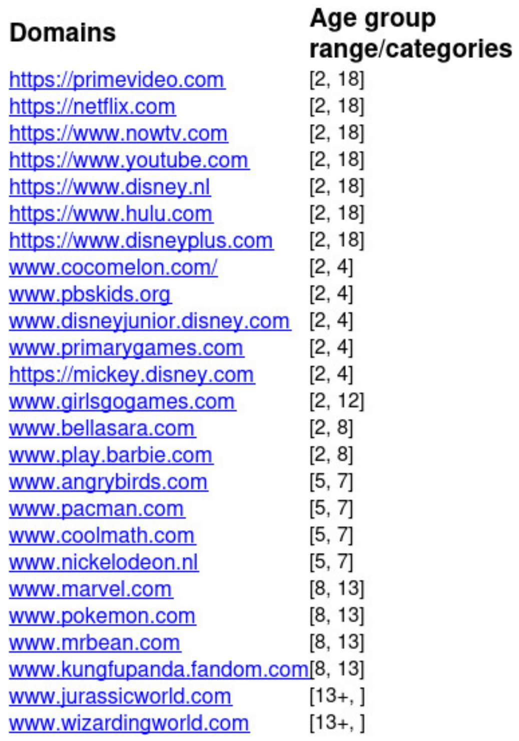


n	ñ
n	ṇ
n	r
n	u
n	η
n	И
n	и
h	и
o	0
o	Ó
o	Ò
o	Ô
o	Ö
o	Õ
o	Ø
o	Φ
p	q
p	ρ
r	γ
s	5
s	$
s	z
t	l
t	ț
t	π
t	τ
u	ú
u	ù
u	û
u	ů
u	ü
u	ų
u	ū
u	µ
v	υ
w	ω
w	ш
x	X
x	ẋ
x	χ
y	ч
y	ý
Y	Ý
y	ý
y	ÿ
y	y
y	λ
z	2

Fig. 13. Confusion table data part - II

Appendix B
List of children domains of various ages

List of children domains of various ages picked based on the methodology mentioned



Domains	Age group range/categories
https://primevideo.com	[2, 18]
https://netflix.com	[2, 18]
https://www.nowtv.com	[2, 18]
https://www.youtube.com	[2, 18]
https://www.disney.nl	[2, 18]
https://www.hulu.com	[2, 18]
https://www.disneyplus.com	[2, 18]
www.cocomelon.com/	[2, 4]
www.pbskids.org	[2, 4]
www.disneyjunior.disney.com	[2, 4]
www.primarygames.com	[2, 4]
https://mickey.disney.com	[2, 4]
www.girlsgogames.com	[2, 12]
www.bellasara.com	[2, 8]
www.play.barbie.com	[2, 8]
www.angrybirds.com	[5, 7]
www.pacman.com	[5, 7]
www.coolmath.com	[5, 7]
www.nickelodeon.nl	[5, 7]
www.marvel.com	[8, 13]
www.pokemon.com	[8, 13]
www.mrbean.com	[8, 13]
www.kungfupanda.fandom.com	[8, 13]
www.jurassicworld.com	[13+,]
www.wizardingworld.com	[13+,]

Fig. 14. children domains of various ages

Appendix C
Protective domain registration and their blacklisting count

TABLE XI
WHOIS query on domains and blacklisting number

Domain	WHOIS registrar matching count	Blacklisting count
youtube	36	2
primevideo	16	0
netflix	54	4
nickelodeon	0	0
jurassicworld	1	0
mickey.disney	1	0
pacman	0	0
pokemon	10	1
disneyjunior.disney	1	0
nowtv	2	0
pbskids	3	0
marvel	8	0
kungfupanda.fandom	3	0
cocomelon	7	0
girlsgogames	5	0
mrbean	13	0
angrybirds	5	0
disney	0	0
play.barbie	3	0
wizardingworld	0	0
bellasara	4	0
hulu	11	0
primarygames	9	0
coolmathgames	4	0
disneyplus	9	0