Third Party Application

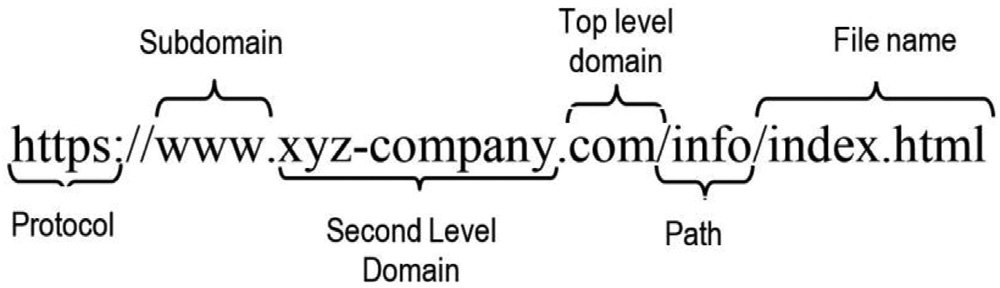
Attackers use different types of techniques for not to be de- tected either by security mechanisms or system admins. In this section, some of these techniques will be detailed. To understand the approach of attackers, ﬁrstly, the components of URLs should be known.

In the standard form, a URL starts with its protocol name used to access the web page. After that, the subdomain and the Sec- ond Level Domain (SLD) name, which commonly refers to the or- ganization name in the server hosting, is located and ﬁnally the Top-Level Domain (TLD) name, which shows the domains in the DNS root zone of the Internet takes place. The previous parts com- pose the domain name (host name) of the web page; however, the inner address is represented by the path of the page in the server and with the name of the page in the HTML form.

Although SLD name generally shows the type of activity or company name, an attacker can easily ﬁnd or buy it for phishing. The name of SLD can only be set once, at the beginning. However, an unlimited number of URLs can be generated by an attacker with extending the SLD by path and ﬁle names, because the inner ad- dress design directly depends on attackers.

The unique (and critical) part of a URL is the composition of SLD and TLD, which is named as domain name. Therefore, cyber- security companies make a great effort to identify the fraudulent domains by name, which are used for phishing attacks. If a do- main name is identiﬁed as phishing, the IP address can be easily blocked to prevent from accessing the web pages located in it.

To increase the performance of the attack and steal more sen- sitive information, an attacker mainly uses some important meth- ods to increase the vulnerability of victims such as the use of ran- dom characters, combined word usage, cybersquatting, typosquat- ting, etc. Therefore, detection mechanisms should take into consid- eration of these attack methods.

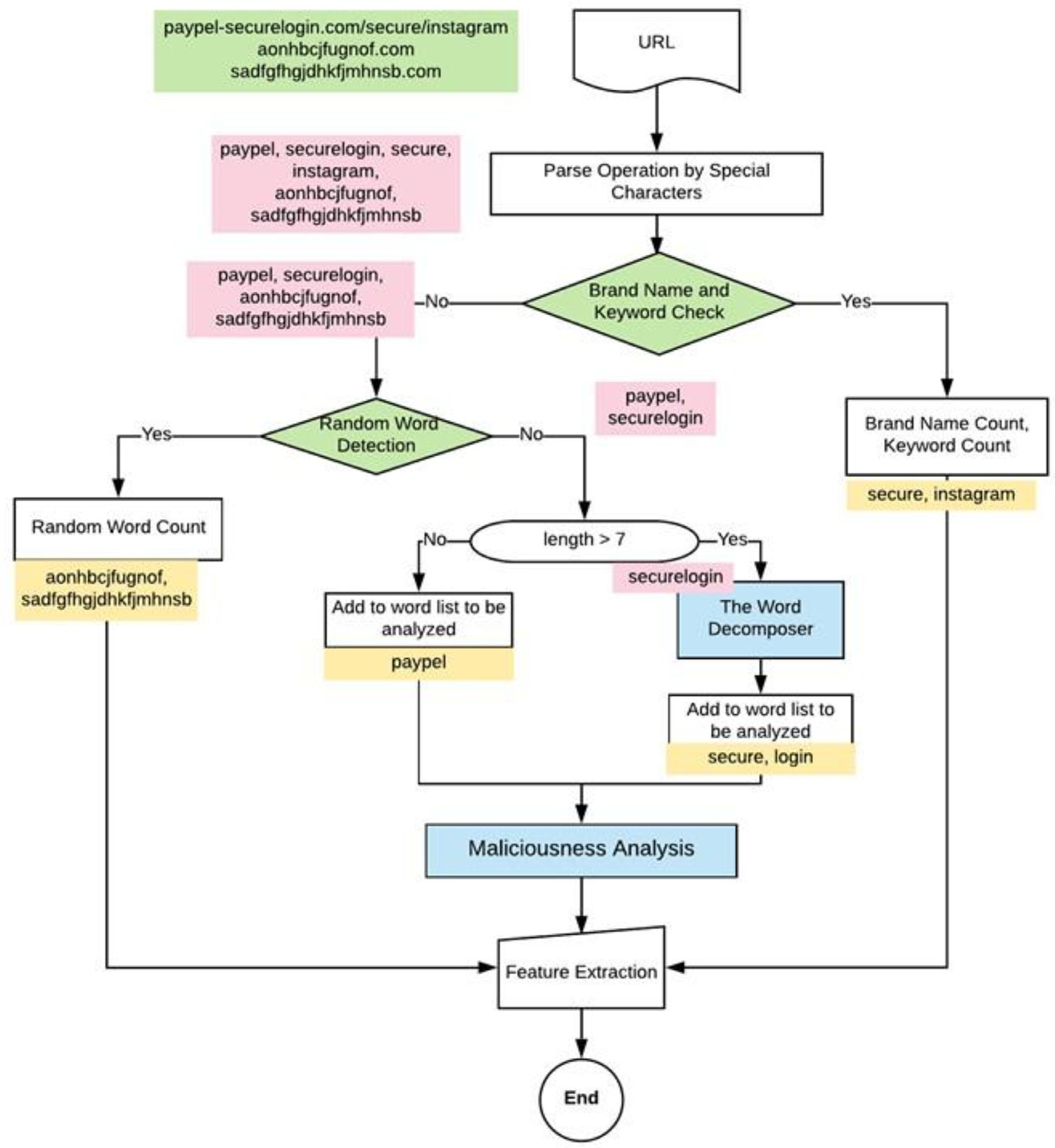


**Data Pre-processing**

A URL consists of some meaningful or meaningless words and some special characters, which separate some important components of the address. For instance, a dot mark is used for separating the SLD and TLD. Similarly, the domain names and the subdomain names are also separated by the same character. However, in the path address folders are separated by the sign. Additionally, each component of the URL may also contain some separation marks such as “.”, “\_” etc. as can be seen in the fol- lowing example “xyz\_company.com”. Similar characters can also be used with “=”, “?”, “&”, characters in the ﬁle path area. Therefore, in the data preprocessing part, ﬁrstly, each word is extracted from the URL and then they are added to the “word\_list” to be ana- lyzed in the ongoing execution. Additionally, the similarity or these words with the most targeted websites and random created words are also detected in this module.

The main aims of data pre-processing part are as follows:

1. Detecting the words, which are similar to known brand names,
2. Detecting the keywords in the URL



Phishing Detection Model:

