Antiphishing Application using Web Document Analysis

Taylor, Onate Egerton & Emu, Jerimiah
Rivers State University,
Port Harcourt, Nigeria.
tayonate@yahoo.com, JEMU@tenaris.com

Abstract
Major security issues for banking and financial institutions are Phishing. Phishing is a webpage attack, it pretends a customer web services using tactics and mimics from unauthorized persons or organization. The false e-mails often look surprisingly legitimate and even the Web pages where users are asked to enter their information may look real. Phishing is similar to fishing in a lake, but instead of trying to capture fish, phishers attempt to steal personal information. This paper gives brief information about phishing, its attacks, steps that users can take to safeguard their confidential information. This paper also shows a survey conducted by net craft on phishing.

Keywords: *anti-phishing technologies, identity theft, Network security, Phishing attacks*

I. Introduction

The World Wide Web has become a huge information store that is growing at a rapid rate, both in the number of web sites and in the volume of the information available, making it now the largest information and knowledge repository. Such vast information urges computer-based efficient and reliable information processing techniques. However, the Web is quite heterogeneous and the information on the Web is organized in a pretty chaotic way. Moreover, most webpages are encoded in the Hyperlink Markup Language (HTML), which is designed mainly for content presentation and suffers poor capability for semantic expression. It is not easy for computer to understand the web documents and thus process them intelligently. The most common purpose of phishing scams includes:

- **Theft of login credentials** – typically credentials for accessing online services such as eBay, Hotmail, etc. More recently, the increase in online share trading services has meant that a customer's trading credentials provide an easy route for international money transfers.
- Theft of banking credentials – typically the online login credentials of popular high-street banking organizations and subsequent access to funds ready for transfer.
- Observation of Credit Card details – access to a steady stream of credit card details (i.e. card number, expiry and issue dates, cardholder's name and credit card validation (CCV) number) has immediate value to most criminals.
- Capture of address and other personal information – any personal information, particularly address information, is a highly saleable and in constant demand by direct marketing companies.
- Distribution of botnet and DDoS agents – criminals use phishing scams to install special bot and DDoS agents on unsuspecting computers and add them to their distributed networks. These agents can be rented to other criminals.

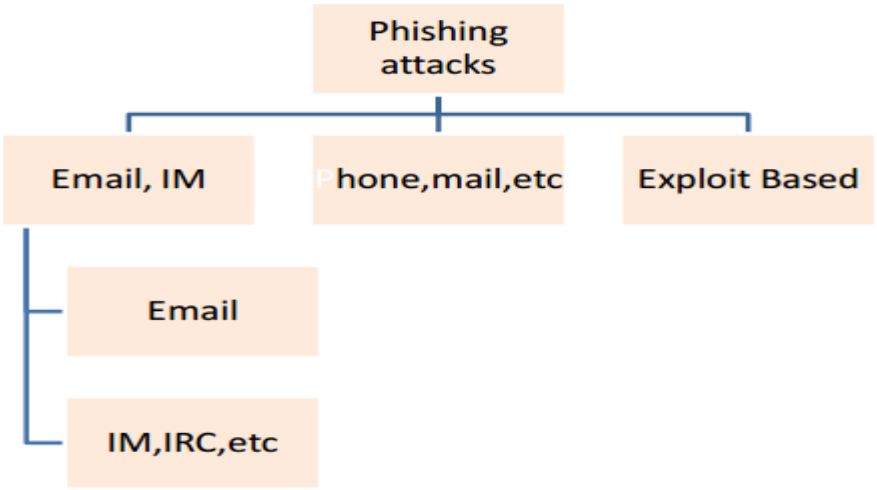

Fig1. Phishing attacks

Attack Propagation – Through a mixture of spear phishing and bot agent installations, phishers can use a single compromised host as an internal "jump point" within the organization for future attack. The proposed phishing website detection system will detect threats and indicate that e-mails, websites or the URL's are not secured and help the user avoid the hacker's trap. Such a type of detection builds confidence in both the users and the Internet community. The phishing website detection system will guide users by providing knowledge of Internet threats. In phishing detection, there are two types of techniques: the white list technique and the heuristic based mechanism. These two techniques act as filters in detecting phishing websites. In white list technique, a few anti-phishing websites are listed. If the user accessed websites are not in the white list, then these will be concluded as phishing websites. The heuristic based mechanism works with various aspects like keywords and domain name to decide whether the website is a phishing website or not. The rest of the paper is as follows: Section II discusses about the background, section III presents the design and implementation of the system, section IV describes the evaluation procedure and results and final conclusions are made in section V.

II. Related Work

Phishing solutions can be broadly classified into four categories [11]. They are:

Malicious Web site URLs:

An approach to this problem based on automated URL classification, using statistical methods to discover the telltale lexical and host-based properties of malicious Web site URLs. These methods are able to learn highly predictive models by extracting and automatically analyzing tens of thousands of features potentially indicative of suspicious URLs.

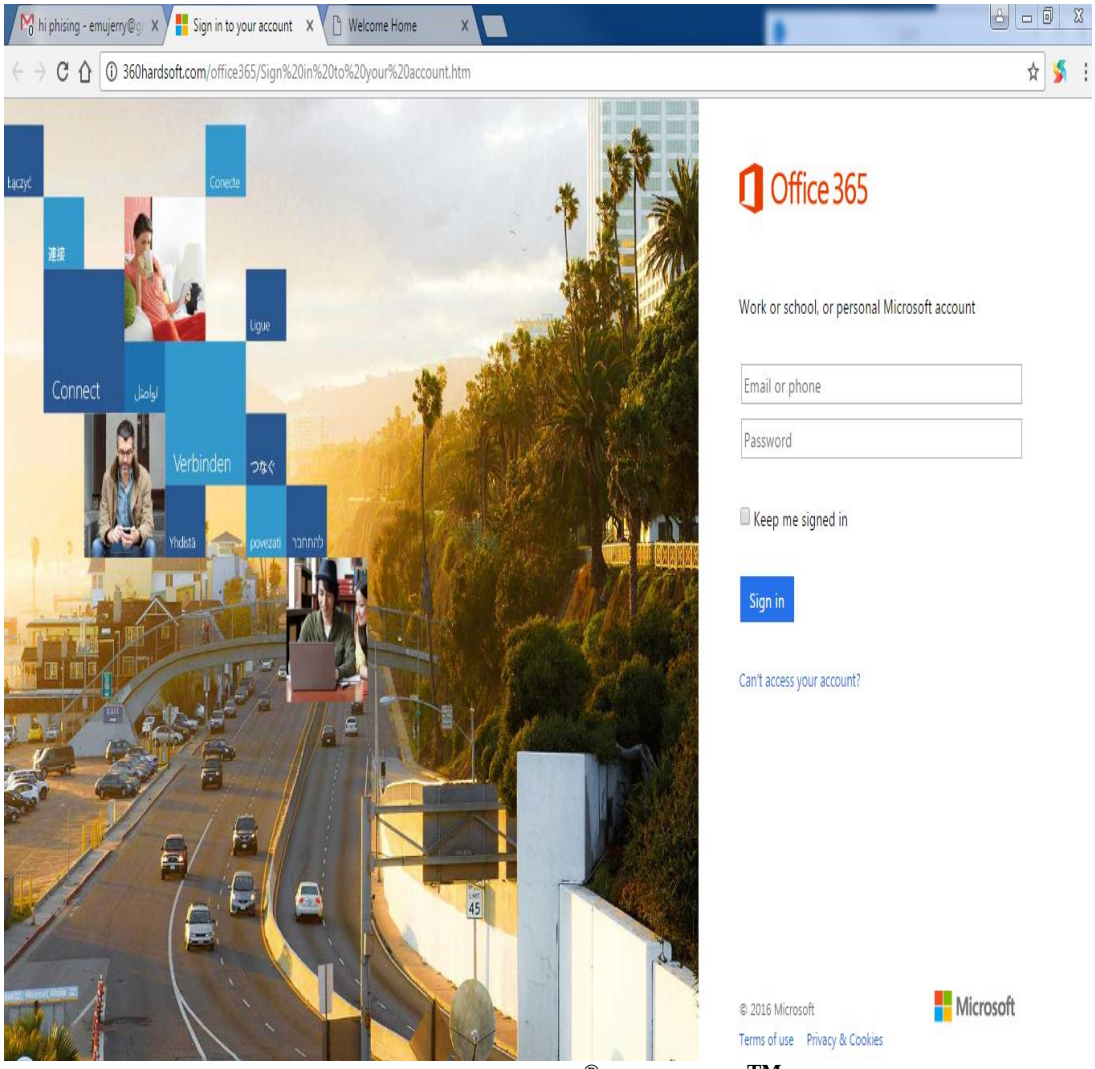

Figure 2: Phishing webpage attack on Microsoft® Office 365™ - office.com

Page Rank:

This work uses the PageRank value and other features to classify phishing sites from normal sites. We have collected a dataset of 100 phishing sites and 100 legitimate sites for our use. By using this Google PageRank technique 98% of the sites are correctly classified, showing only 0.02 false positive rates and 0.02 false negative rates.

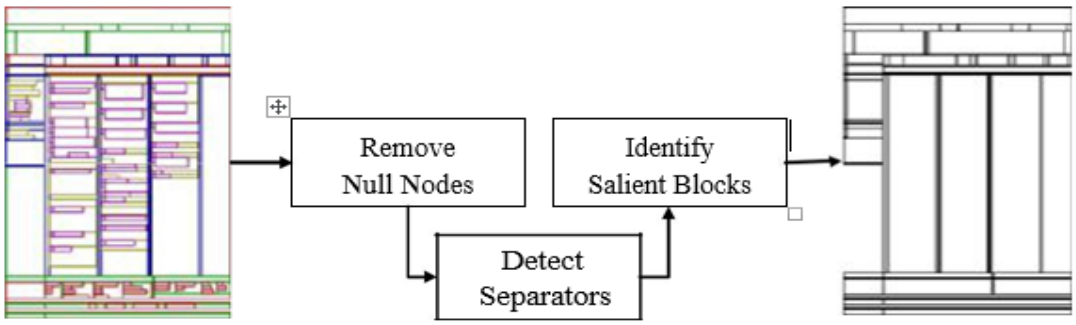

Fig.3: Salient block decomposition (Different colors represent different levels in the DOM tree)

CANTINA:
A novel content-based approach for detecting phishing web sites. CANTINA takes Robust Hyperlinks, an idea for overcoming page not found problems using the well-known Term Frequency / Inverse Document Frequency (TF-IDF) algorithm, and applies it to anti-phishing. We described our implementation of CANTINA, and discussed some simple heuristics that can be applied to reduce false positives. We also presented an evaluation of CANTINA, showing that the pure TF-IDF approach can catch about 97% phishing sites with about 6% false positives, and after combining some simple heuristics we are able to catch about 90% of phishing sites with only 1% false positives

Behavior based Detection:
A novel approach to detect phishing websites based on analysis of users' online behaviors – i.e., the websites users have visited, and the data users have submitted to those websites. Such user behaviors cannot be manipulated freely by attackers; detection based on those data can not only achieve high accuracy, but also is fundamentally resilient against changing deception methods [9].

Lexical Analysis
This paper presents a lexical URL analysis (LUA) technique to enhance the classification accuracy of anti-phishing email filters. Although the LUA feature is primarily focused to classify phishing websites, it proved to be effective to classify email messages due to the fact that most phishing email messages contain URLs. According to the performance evaluation, the LUA feature proved to be effective in enhancing the classifier's accuracy in all features subsets

Detecting Webpage Source Code
We propose a phishing detection approach based on checking the webpage source code, we extract some phishing characteristics out of the W3C standards to evaluate the security of the websites, and check each character in the webpage source code, if we find a phishing character, and we will decrease from the initial secure weight. Finally, we calculate the security percentage based on the final weight, the high percentage indicates secure website and others indicates the website is most likely to be a phishing website. We check two webpage source codes for legitimate and phishing websites and compare the security percentages between them, we find the phishing website is less security percentage than the legitimate website; our approach can detect the phishing website based on checking phishing characteristics in the webpage source code.

Behavior based Detection:
A novel approach to detect phishing websites based on analysis of users' online behaviors – i.e., the websites users have visited, and the data users have submitted to those websites. Such user behaviors cannot be manipulated freely by attackers; detection based on those data can not only achieve high accuracy, but also is fundamentally resilient against changing deception methods

III. Evaluation Procedure and Algorithm
Phase I*:* Blacklist:
When user enters into the web browser and type a URL in web page. Check whether the site is phishing or not in the black listing. It is holding a phishing URLs in the list. If any illegitimate site will appear, it will alert user web browser. Otherwise it goes to web parsing and heuristics terms.

Fig. 4: page segmentation

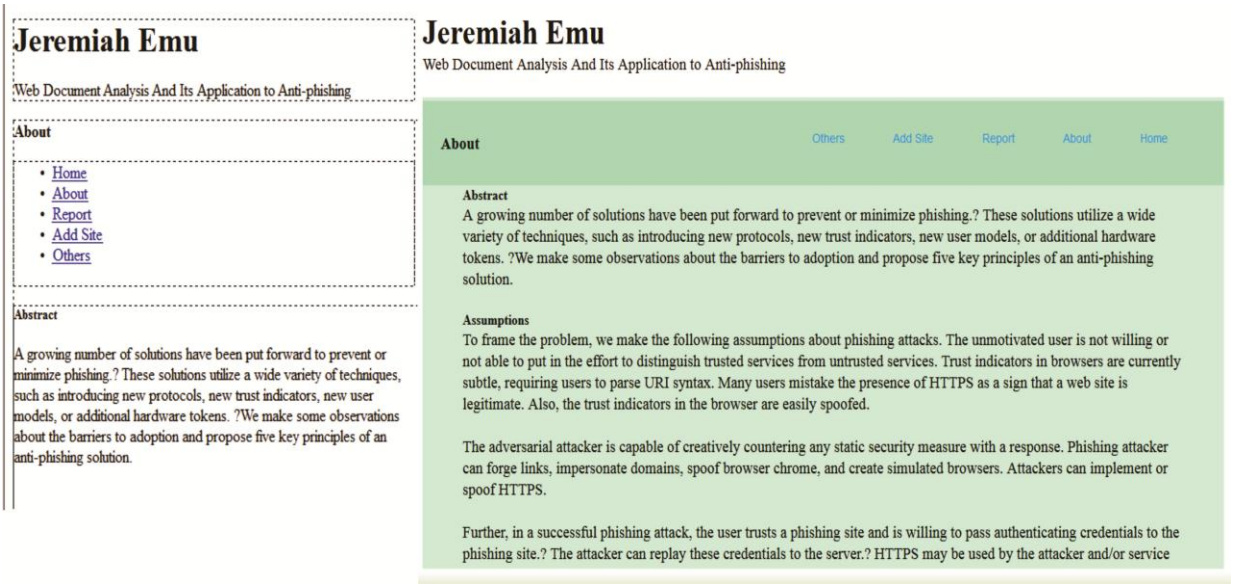

Fig. 5: **Page response**

Phase II: **Scripting in the source code:**
A normal web user does not have knowledge whether a website is a malware. In the following steps are;

a) Web parsing:
Web parsing is a process in which every HTML code from the source of the web page is parsed. Tags such as <>, html, br, textbox, regular expressions, etc., will be eliminated in this method each and every HTML tag in the source of the webpage are parsed.

b) Separating the Required Tokens:
After parsing is done on the source of the webpage only the data and information other than the unwanted links and tags will be displayed. After parsing the web page, the required tokens are separated. A token could be a keyword, an operator, or a punctuation mark.

c) Classification of Scripting Tokens

If any external tokens are found while parsing, must be classified. These external tokens are created by hackers generally known as man-in-the-middle. Finally, we text identification from the scripting and weight based find out phish site or legitimate site.

Phase III: **Classification of Heuristics:**

In this phase classification a url by using heuristics based. We refer before; finally obtain a phishing or legitimate site. The contributions of this paper are: 1) to show how PageRank value can be useful to detect phishing. 2) An implementation to show high accuracy of classification of phished web sites. 3) Considering other features like age of the domain, suspicious URL, whether the domain contains IP address or not, number of dots and whether it is taking user personal information as input or not

Our major contributions are:

1. An anti-phishing strategy that can be used as an enterprise solution by the legitimate webpage owner to take the initiative to detect phishing/forged webpages which are visually similar to the true webpage.

2. The visual similarity metrics in three aspects: block level similarity, layout similarity, and overall style similarity.

3. The style features and the method for measuring the overall style similarity of webpages.

IV. Conclusion

The purpose of this document is to provide a set of recommendations to the domain registrar community that can substantially reduce the risk and impact of phishing on consumers and business worldwide. The recommendations focus on 3 areas where registrars can be of assistance:

1. Evidence Preservation for Investigative Purposes:

As registrars are in direct contact with the criminals as they are registering fraudulent domains (typically through the registration process on the registrar's website), they may have the ability to acquire key important evidence that can be later used by law enforcement to identify and prosecute the phishers. This document enumerates the type of evidence that can be collected during the domain registration process by the registrar that would be helpful to law enforcement. We encourage registrars to collect and store as much of this evidence as is feasible in their circumstances to achieve the best chances of law enforcement catching the criminals. Increasing the risk for these criminals of capture, prosecution, and incarceration should have a significant deterrence impact and eventually result in a reduction of these types of crimes.

2. Proactive Fraud Screening:

With a bias towards not impacting legitimate customers, anything that registrars can do to complicate the domain registration process in order to frustrate the phishers and limit their ability to perform fraudulent domain registrations on a large scale is highly beneficial. This document suggests some lightweight processes registrars can put in place to identify fraudulent activity before the domain registration takes effect. The harder it is for the criminals to commit these frauds, the more likely it is that they will move off to something else that requires less effort on their part.

3. Phishing Domain Takedown:

Once a phishing site goes live and is promoted by the phisher, it is imperative that it be taken down as quickly as possible in order to limit its impact and the number of potential victims. As part of the takedown process, anti- phishing organizations typically contact both the hosting provider of the phishing website, as well as the registrar or registry responsible for a fraudulent domain registration. Document contains best practices that registrars can use to process the takedown requests in the most optimized fashion and limit the victim's financial losses.

V. References

Adeniji, A. A. (2009). Cost Accounting, A Managerial Approach. Lagos State, Nigeria: El- Toda Ventures limited publishers.

Adeniyi A. A. (2014). An insight into management accounting. 8th edition. Value analysis publishers lagos.

Ado R. (2007). Exploring the facts of ABC and its Practicability in the Nigerian Manufacturing Industries, *The New Breed Accountant, BUK pg 10-16.*

Ali U. (2010). Cost and Management Accounting Practices: A Survey of Manufacturing Companies . *Eurasian Journal of Business and Economics 2010, 3 (6), 113-125.*

Ama, G.A.N (2001). Management and cost accounting: current theory and Practice. Abia. Amazons publishers ventures.

Ammar, A., Awad S. H., Eric V. N., and Jeffrey S. R. (2003). Indicators variables model of firm's size-profitability relationship of electrical Contractors using financial and economic data. *Journal of Construction Engineering and Management 129 (March): 192-197.*

Asika N. (2008): Research Methodology in the Behavioral science. Nigeria. Longman Nig Plc

Ax, C., Greeve, J., & Nilsson, U. (2008). The Impact of Competition and Uncertainty on The Adoption of Target Costing. *International Journal of Production Economics, 115(1), 92-103.*

Babalola, Y. A. (2013). The Effect of Firm Size on Firms Profitability in Nigeria. *Journal of Economics and Sustainable Development ISSN 2222-1700 (Paper) ISSN 2222-2855 (Online) Vol.4, No.5, 2013*. www.iiste.org.

Bhayani S.J., (2010) "Determinant of Profitability in Indian Cement Industry: An Economic Analysis", *South Asian Journal of Management, 17 (4), pp. 6-20.*

Bhimani A. and Pigott .D. (1992). Implementing ABC: A Case Study of Organizational and Behavioural Consequences, Management Accounting Research, Vol., 3: 119-132.

Bonzemba, E. L., & Okano, H. (1998). The Effects of Target Costing Implementation on an organizational culture in France. Osaka: Osaka City University.

Chartered Institute of Management Accountants (CIMA) (2001). Activity Based Management: An Overview, London: 63 Port Land place.

Dabor E.L, & Erhgbhe E. (2005). Implementing ABC in the Health Service Industry: A Survey of Some Private Clinics in Benin City, *Nigeria, Bayero University Journal of Accounting Research, Vol.1, No. 2 Pg 56-67.*

Dalton, D. R, Daily, C. M, Johnson, J. L & Ellstrand, A. E (1999). Number of directors and financial performance: A meta-analysis, *Academy of Management Journal*, 42 (6), 674-686.

Drury C. (2002). Management and Cost Accounting, Book Power/ELST, London: 5th edition.

Drury, C. (2005). Management and Cost Accounting (6th ed.). London: Thomson Learning

Faraway, R. E. (1984).Strategic management: A stakeholder approach, Prentice-Hall, Englewood Cliffs, NJ.

Eyisi, S. A. (2009). Cost Accounting: Theories and Practice. Enugu. Ayi- best Publishers.

Eze, J.C and Ani, W.U. (2009). Intermediate cost accounting. Enugu: JTC publishers.

Ezeamama, M.C (2010). Fundamentals of financial management: a practical guide. Enugu. Ema press limited.

Fridh, G., & Borgernas, H. (2003). The use of target costing in Swedish manufacturing firms. Goteborg University: School of Economics and commercial law.

Appendix A: Code and Segment

The features for each block are extracted in the feature extraction step in the corresponding webpage processing module and form the feature set of the block.

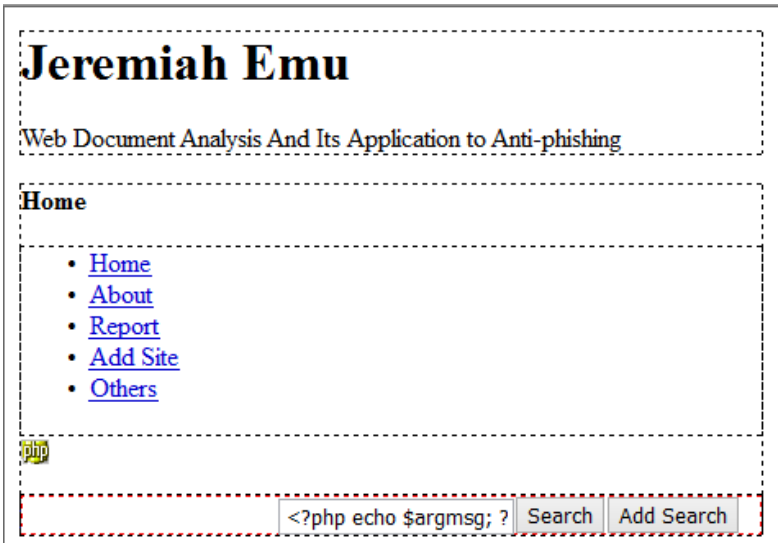

Figure 6: Part of the true "http://360hardsoft.com/login" webpage and its segmentation result.

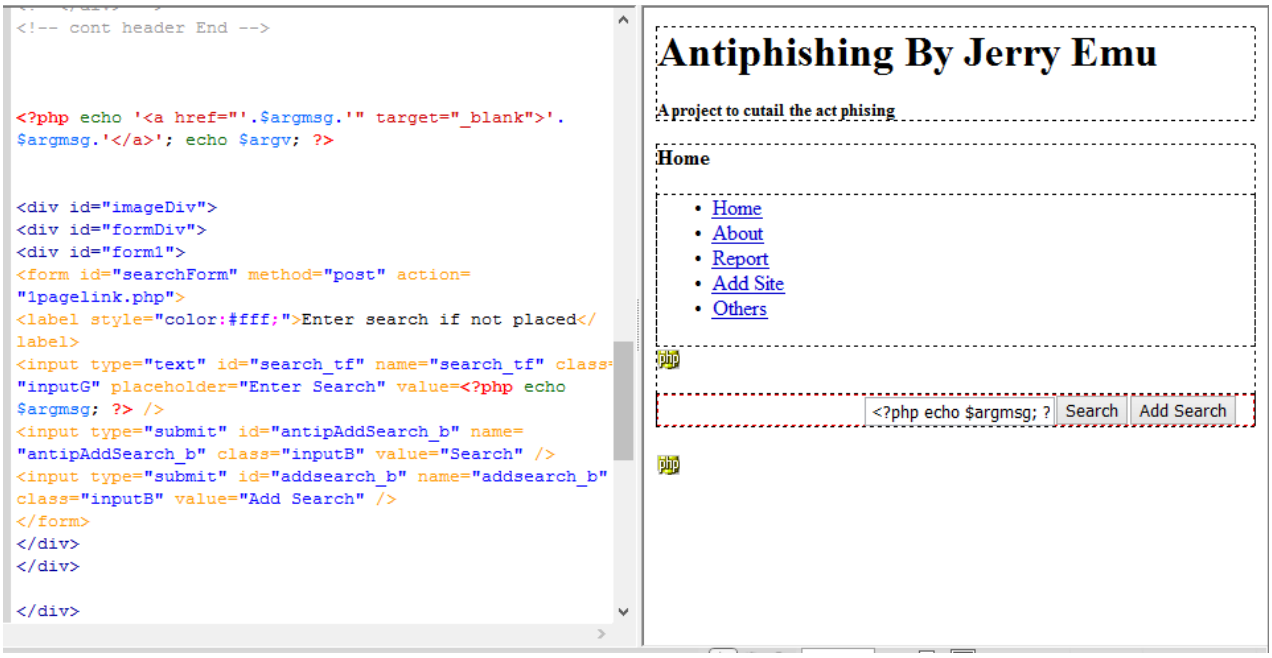

Figure 7: Part of a phishing webpage of the true "http://360hardsoft.com/login" webpage