Scientific Journal of Impact Factor (SJIF): 4.14

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 3, Issue 5, May -2016

# Analysis of Decisive and Non-Intrusive Technique to Combat Form Spam

Parita Chandan[1], Chintan Thacker[2], Manish Saxena[3]

[1,2]Department of Computer Engineering, Gujarat Technological University,
[3]MCA Department,Feroze Gandhi Institute of Engineering and Technology

**Abstract** — *Spammers and the spam e-mails are causing huge losses to businesses & individuals on a regular basis in terms of time & money. Spam filtration is an automated technique to identify SPAM and HAM. Various types of spam filters are designed with different approaches, each having their own pros and cons. But after studying various research works it was found that among all, the two most widely used methods are HONEYPOT and CAPTCHA. These two methods are also available in lots of variants. Through this research work, I have tried to analyse and identify best from the above two methods to combat form spam. At first, I studied HONEYPOT and CAPTCHA. Then, I implemented these two methods on various forms. Later, compared and contrasted these two methods on certain parameters, most importantly on false positive (means the test said the message was spam, when in reality it wasn't.) base and tried to find the method which will be able to identify more spams in less time, which is more users friendly, provide easy accessibility easily manageable and, most importantly minimize the spam generated from Contact and Feedback forms on public and social networking CMS websites.*

***Keywords-*** *Honeypots, Spambots, CAPTCHA, SPAM and HAM.*

## I. INTRODUCTION

The word spam means junk mails. The unsolicited emails that are received by any person in his / her mailbox are called spam. These junk mails are usually sent in bulk for advertising and marketing some products. Lots of space in your mailbox is occupied by these junk mails. Sometimes it eats up your valuable space so that the genuine mails are bounced back to the sender if the whole lot of space is occupied by the junk mails. Hence there comes a necessity to filter out those junk mails from your mailbox[15].

**Incentives to spam include:**

- Advertising on a massive scale;
- Manipulating online voting systems;
- Destabilizing a critical human equilibrium (i.e. creating an unfair advantage);
- Vandalizing or destroying the integrity of a website;
- Creating unnatural, unethical links to boost search engine rankings;
- Accessing private information;
- Spreading malicious code.


*Figure 1. Spammy Ads*

 

Those who run websites know that this is a big business and a big problem. On one side of the coin is the spammer; on the other is the humble website owner, who experiences common problems that include:

- Blogs and forums that sink under the weight of spam posts,
- Accounts that are registered under false pretences for unlawful purposes,
- Bots that ruin the dynamics of a website,
- A dive in the quality of content and the user experience.

For those who operate their own websites and allow users to register for an account to be able to leave comments or to sign up for services or post in forums, one of the fights problems spammers. spam is a huge problem for site owners. It can cost businesses time and money .To fight spam many sites put CAPTCHA on their forms, This CAPTCHA can stop spambots from spamming, but they can also stop users from filling out forms. since users are made to type random words that don't make sense- not only that, but the letters are so wrapped and distorted that they are hard for anyone to read. Users often have to try CAPTCHA many times to get it right, that's why most users avoid filling out the form when they see one .It is good that CAPTCHA stops spams but sometimes it comes with the cost of losing the legitimate user[12].

CAPTCHA is not the ultimate solution for user authentication, as it can be easily compromised by online users of CAPTCHA solving companies. So we need a solution for such companies. The CAPTCHA challenge is not a challenge any more with budding solutions for automatic CAPTCHA solving. Bypass CAPTCHA (www.bypasscaptcha.com), CAPTCHA Brotherhood (www.captchabrotherhood.com), Image Decoders (www.imagedecoders.com), Imagetyperz (www.imagetyperz.com), are some of the examples of the CAPTCHA solving services. They are not using any image recognition techniques but they have deployed humans to solve these CAPTCHA puzzles. With their economical working model,- they have plenty of online users to solve the CAPTCHA puzzles 24 X 7. Various CAPTCHA methods have been defined till now, but all come with some problems. So through this research, I wish to give the best method that will fight spam without frustrating users[1].

The Anti-Spam strategies can be categorized as follows:

Detection based anti-spam strategies attempt to identify spam and remove it or reduce its prominence; whereas Demotion based anti-spam strategies attempts to lower the ranking of spam in ordered lists; and Prevention based strategies attempts to make contribution of spam more difficult by changing interfaces or limiting user actions [1].

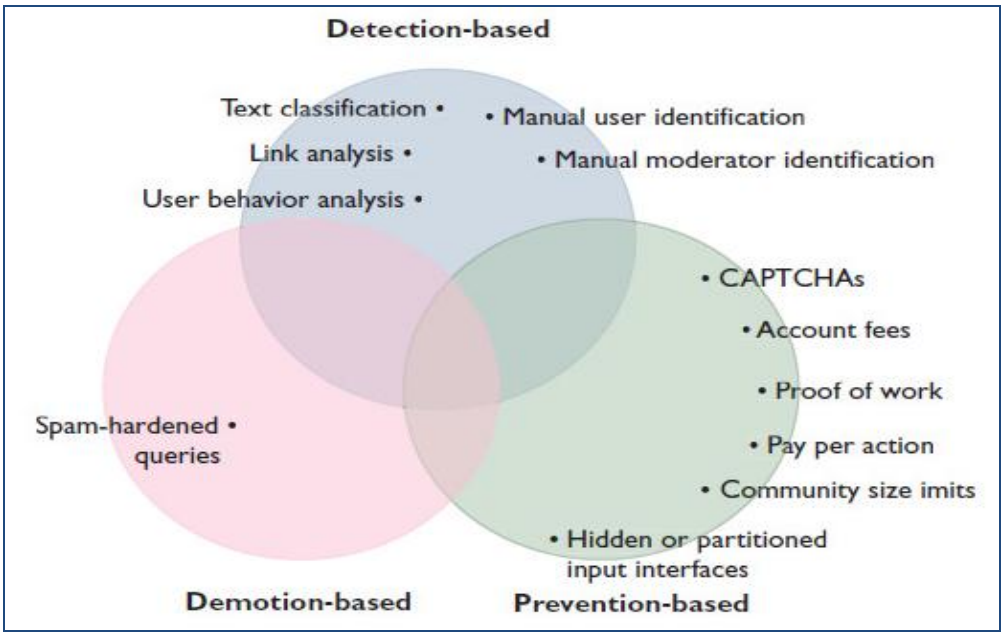


*Figure 2. Anti-Spam strategies*

## II. SCOPE OF RESEARCH

Objective of this research work is to find the most efficient and user friendly spam filter technique which identify more false positive, which is more users friendly, provide easy accessibility, easily manageable and, most importantly minimize the spam generated from Contact and Feedback forms on public and social networking CMS websites. This research work is carried out in two phases.


*Figure 3. Spam Filtering Technique*

Phase 1: Study of all the available spamizer techniques from the available research work done by previous researchers and finding the two widely use spamizer methods.
Phase 2: Implementing the two methods and comparing and contrasting these two methods on certain parameters, most importantly on false positive ratio and ability to minimize the spam generated from Contact and Feedback forms too.

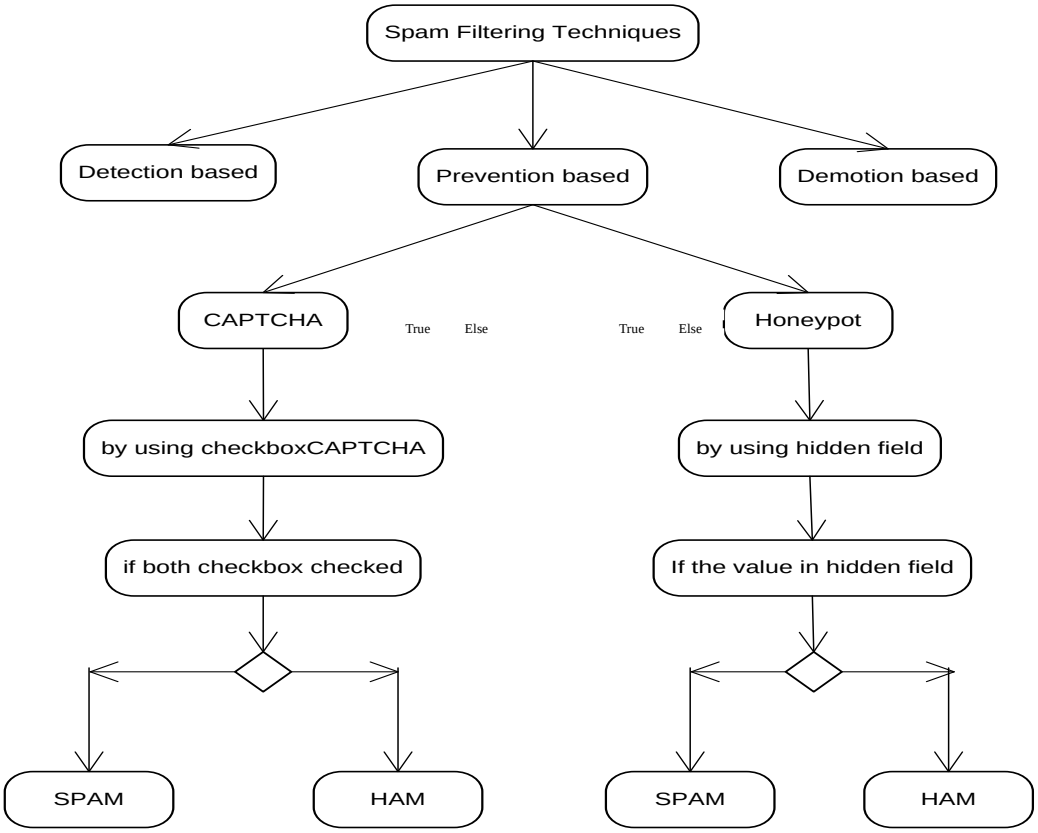


*Figure 4. Activity diagram for CAPTCHA vs Honeypot*

 

## III. COMPARISON OF DIFFERENT TYPES OF SPAM FILTERING TECHNIQUES

*Table 1. Comparison of different types of spam filtering techniques*

| Sr. No | Spam Filtering Techniques | Advantages | Drawbacks |
|---|---|---|---|
| 1 | Checkbox CAPTCHA | Easy Interaction | Confuses a user who doesn't know what a spambot is. |
| 2 | Slider CAPTCHA | Easy Interaction for legitimate user | Doesn't cover accessibility issues. |
| 3 | Puzzle CAPTCHA | Spammers cannot easily fill the puzzle | Only works for limited no. of users on a system. |
| 4 | Socio CAPTCHA (SCAP) | Full of fun & surprises for the user. | User should have social media account. User/Account holder has provided enough information while creating profile |
| 5 | Honeypot Method | Legitimate user need not prove himself. | Honeypots can only track activity that interacts with it. |
| 6 | Verified sign-in | Includes the benefit of removing the anonymity. | Invasion of privacy. |
| 7 | Recording User Time Expenditure | Easy to implement | Some sneakier bots are programmed to take longer to fill out forms to specifically avoid this trap. |

## IV. IMPLEMENTATION METHEDOLOGY

*I. Check box CAPTCHA*

The checkbox catpcha can stop some spambots, but not all. What's good about this one is that it's smaller and less intrusive than traditional captchas.The checkbox option works by placing a checkbox on a form which users are asked to select or unselect before submission. All it takes is putting a checkbox generated with client side JavaScript on the form. All the user has to do is tick it. No typing necessary.Spambots won't be able to tick the checkbox because they don't parse client side JavaScript, although it's less intimidating to users, because many people browse with JS disable for security and privacy purposes[11].

It rates highly on list of CAPTCHA alternatives, but the terminology should be easy for users to understand. 'I'm not a spambot' is likely to confuse users who don't know what a spambot is . Simply label 'Select this box before pressing submit', but if user confuse and don't fill the checkbox then result will be false positive. So it's not 100% effective[13].

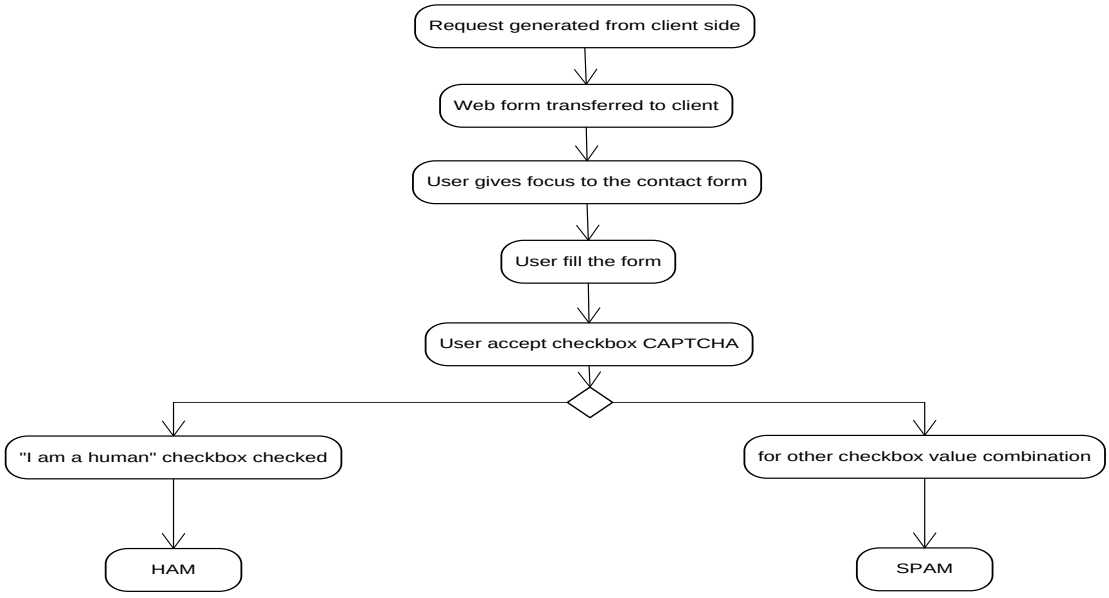


*Figure 5. Activity diagram of Checkbox CAPTCHA*

 

Here, I am using Roboform software to fill the form as spammer.I am using two checkbox without using css/js. If spammer check both the checkboxes. Then counter will be increment and data will not be post and it will be counted as a spam. But if legitimate user check both the checkboxes then result would be a false positive. So we cannot identify that how were real spammers and how were legitimate users who checked both checkboxes by mistake because of lack of understanding or confusion. If the user check only "I am human" checkbox then it will count as HAM and counter will not be incremented and data will be posted.

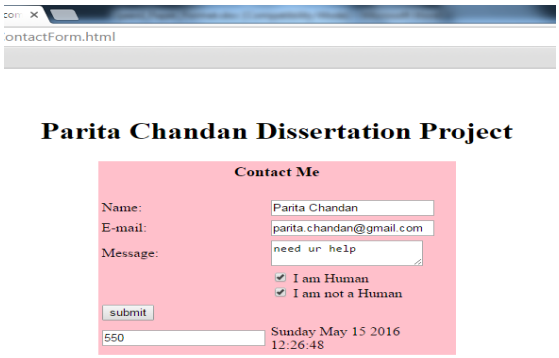


*Figure 5. Snapshot of spammer fill the form using Roboform*

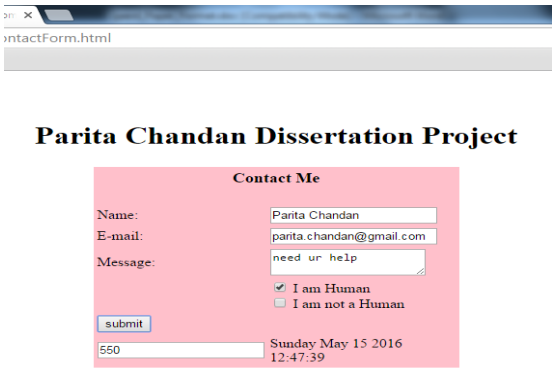


*Figure 6. Snapshot of legitimate user fill the form*

*VII. Recording User Time Expenditure*

Another technical alternative which is hidden from users is the time-based form. The idea behind this is to detect a spambot based on the time it takes to complete a form. Genuine users take a few moments to complete a form, whereas spambots complete forms instantly. Therefore any forms submitted too quickly would be identified as a spambot. We can see this solution working quite well, as long as the time-frame set is practical for users to achieve. In this approach for Spam Detection,we record the 1st Time Stamp when user starts working on a Form, then we are record the final Time Stamp when user submits a Form [1].

Application Time = Submit Time – Initial Time

If the application time is less than or equal to Min. Submit Time while filling the form. Then this method considers it as a spam. Users with cookies enabled, the form may auto-populate, causing the users to be wrongfully identified as a bot so the result will be a false positive.

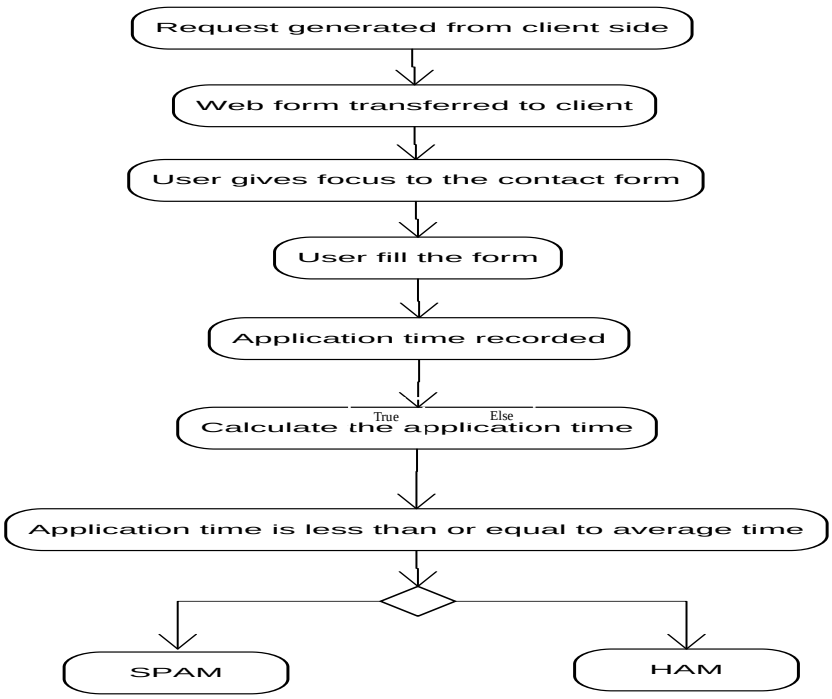


*Figure 7. Activity diagram of recording user time expenditure*

 

Here, I am recording the application time to fill the form by user. User take Min 5sec to fill the form but spammers take Min 1 or 2sec to fill the form. So if application time is less than or equal to Min submit time than it would be counted as a spammer.

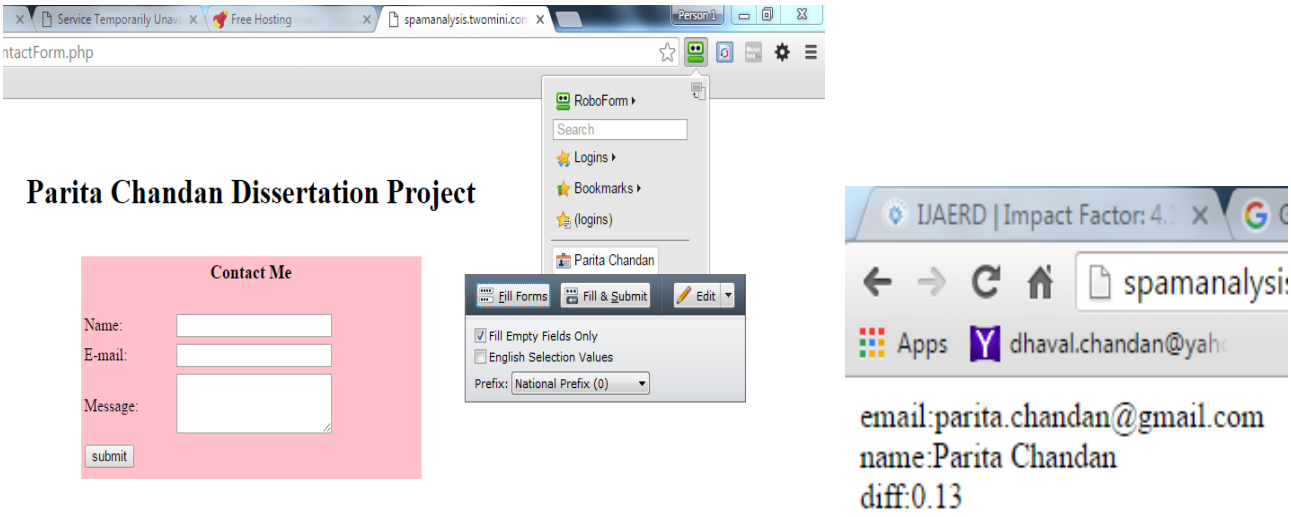


*Figure 8. Snapshot of spammer fill the form using Roboform within second*

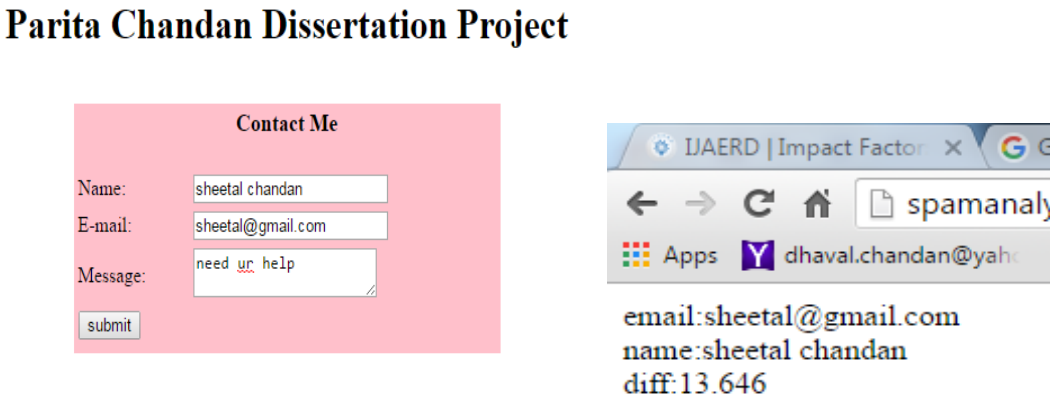


*Figure 9. Snapshot of user fill the form*

*V. HONEYPOT METHOD: Using Hidden Form Field*

Spam-bots always struggle to read CSS or JavaScript on a webpage. One simple solution is to add a completely junk field in every web form, to hide this field using CSS or Java Scripts. For example:

<input type = " text" name="secret" style="display:none;">

Users hate using sites with CAPTCHA. Alternative solutions are available which are not as frustrating as CAPTCHA. The best solutions are those that don't require users to prove they are not spam-bots.Another CAPTCHA which are less intrusive than traditional CAPTCHAs are honeypots. Honeypot are far better than traditional CAPTCHAs[11].

Honeypot CAPTCHAs work by hiding a text field from users through CSS. It'll only accept entries that leave the field blank. Users can't fill out this field because they can't see it. But spambots will see and fill it in. The form will then reject the spambot's entry. This presents accessibility issues for screen reader users who have CSS disabled. If the label on the honeypot field doesn't tell them not fill out the honeypot, they won't know to avoid it. You could give the honeypot field a common label, such as "name", to trick the spambot into filling it in. But it would also trick screen

reader users to fill it in too. They can stop some spambots, but not all. They may also create accessibility issues for some users[1].

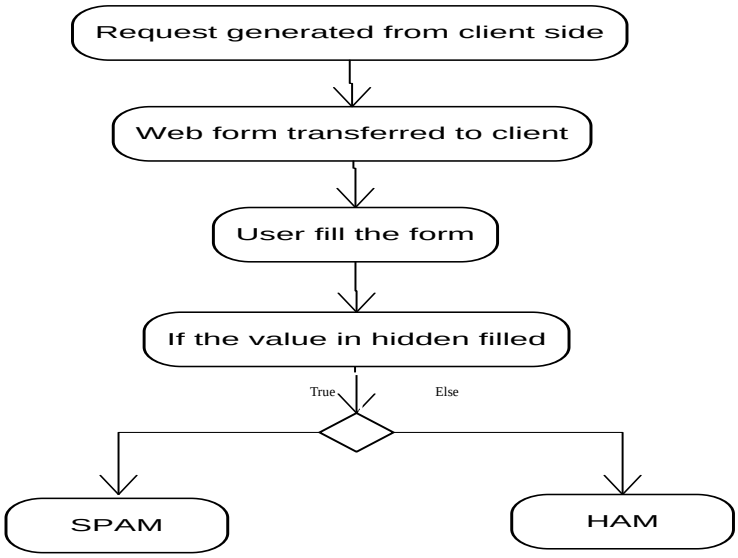


*Figure 10. Activity Diagram for Honeypot*

If we strictly apply condition like only those users can fill the form who have css/js enabled.Then it may overcome accessibility issues, but some Legitimate users will be there who can't fill form, due to css/js being disabled in their browser.

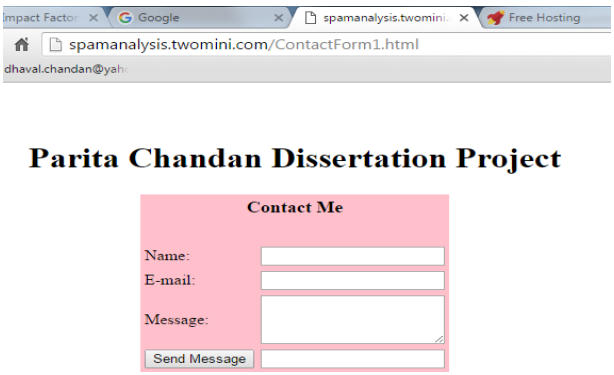


*Figure 11. Snapshot of css enabled form*

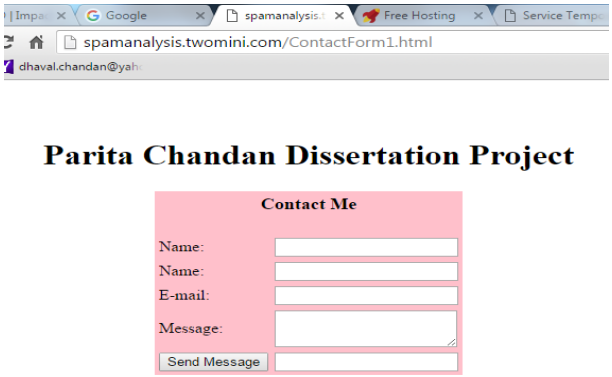


*Figure 12. Snapshot of css Disabled form*

By using text=”hidden” property of the html we can easily identified spambot without using js/css enabled.

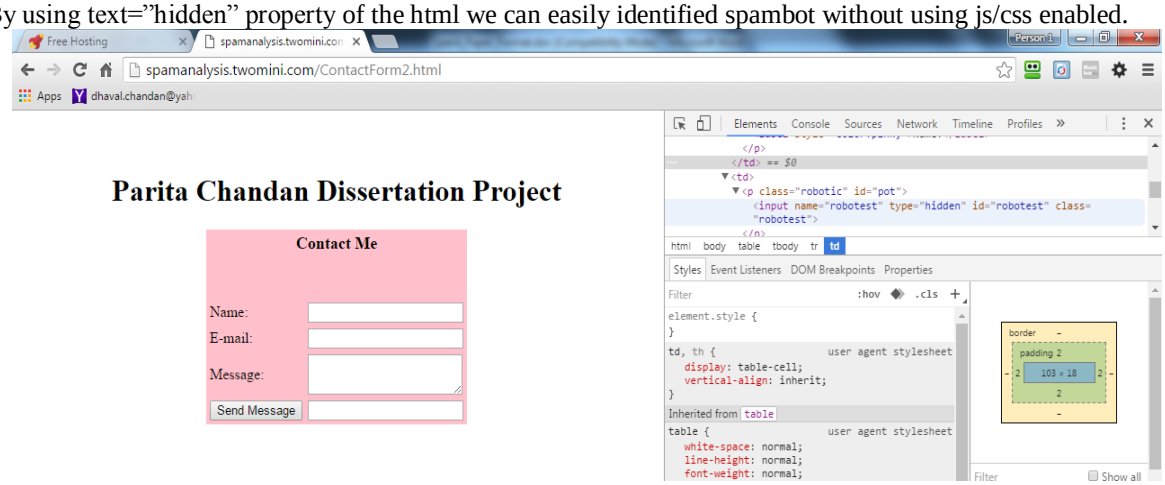


*Figure 13. Snapshot implementation of honeypot by using type=”hidden”.*

 

But there are lots of other ideas:

- Color the fields the same (or very similar to) the background of the page.
- Make an element too small to show the contained honeypot field.
- Use positioning to move a field off of the visible area of the page.

By using same background color, same foreground color, and size of the textfield in pixel we can identified spambot without using js/css enabled.

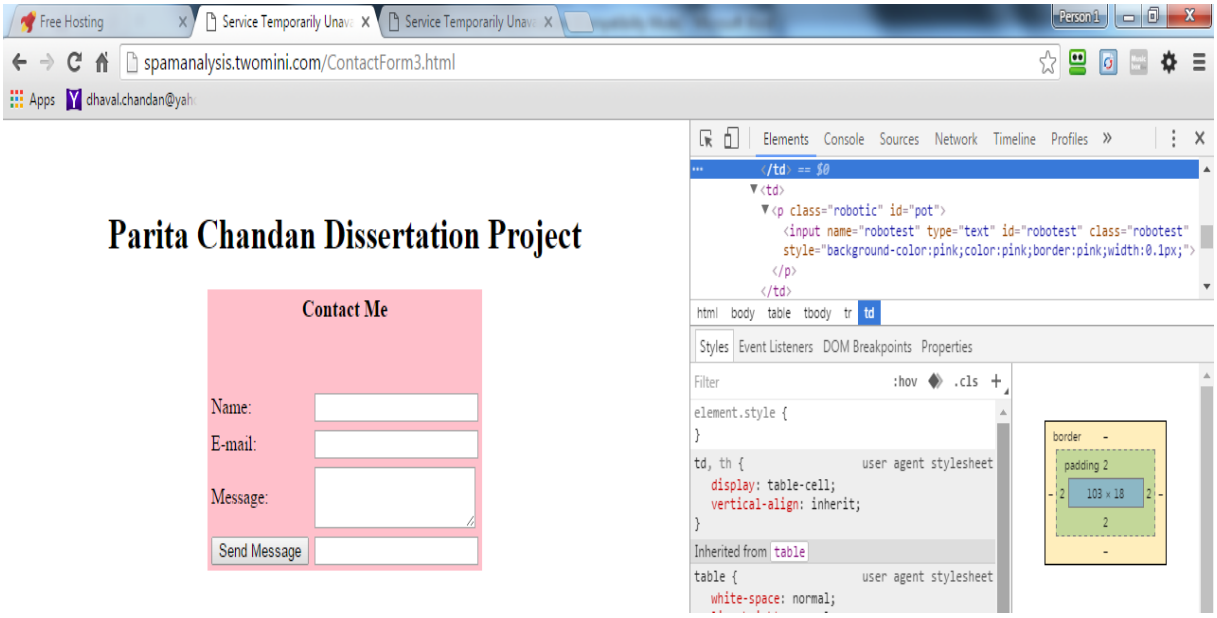


*Figure 14. Snapshot of implementation of honeypot using same background & foreground color*

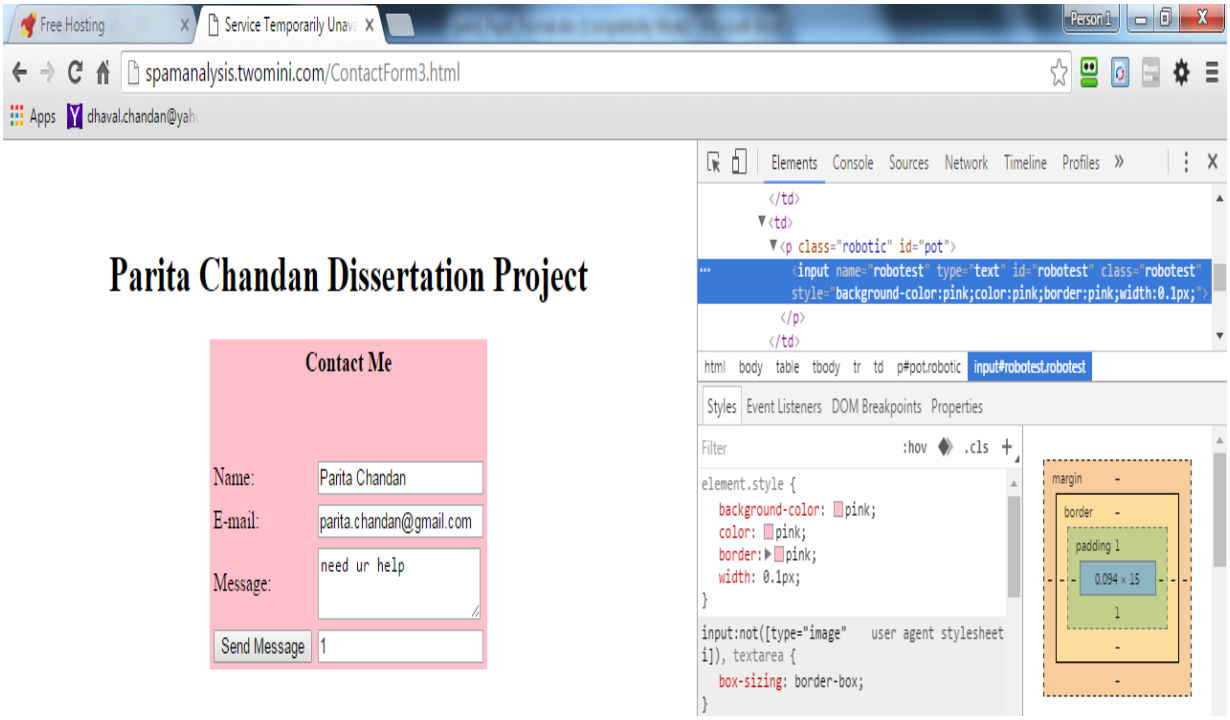


*Figure 15. Snapshot of hidden field fill by roboform*

## V. EXPERIMENTAL RESULT

*Table 2. Experimental Results of filtering techniques*

| Sr. No | Techniques | Total no. of attempts | No of Legitimate users entries | No of Spammers entries | No of SPAM | No of HAM | False positive | Comments |
|---|---|---|---|---|---|---|---|---|
| **1** | Checkbox CAPTCHA | 1000 | 500 | 500 | 550 | 450 | 50 | Confuses users who don't know what a spambot is. |
| **2** | Recording user time expenditure | 1000 | 500 | 500 | 520 | 480 | 20 | User have opted auto-fill property in their browsers |
| **3** | Honeypot by using hidden field in css/js | 1000 | 500 | 500 | 500 | 495 | 0 | 5 Legitimate users who Can't fill form due to css/js disabled |
| **4** | Honeypot by using text="hidden" property of the html | 1000 | 500 | 500 | 500 | 500 | 0 | Without use of js/css |
| **5** | Honeypot by using background & foreground color same of the textfield | 1000 | 500 | 500 | 500 | 500 | 0 | Without use of js/css & text="hidden" property of the html |

The key findings of this experiment are as follows:

- In checkbox CAPTCHA, the number of false positive is 50, because some user checked both the checkboxes.
- Blocking spam by recording user time resulted in 20 false positives. Humans take more time compared to spammers. However, users who have opted auto fill property enabled, got identified as spam as they could fill the form in less than the average time.
- In Honeypot by using hidden field in css/js, The number of false positive is 0. But 5 Legitimate users are there who couldn't fill the form due to css/js being disabled in there browser and it is requirement that js/css is enabled to fill the form.
- In Honeypot by using text="hidden" property of the html the number of false positive is 0. In this method, we are not using js/css to hide a field.
- In honeypot by using background & foreground color same of the text field & text field width set in pixel. The false positive is 0. This method is a simple method without use of js/css & text="hidden" property of the html.

 

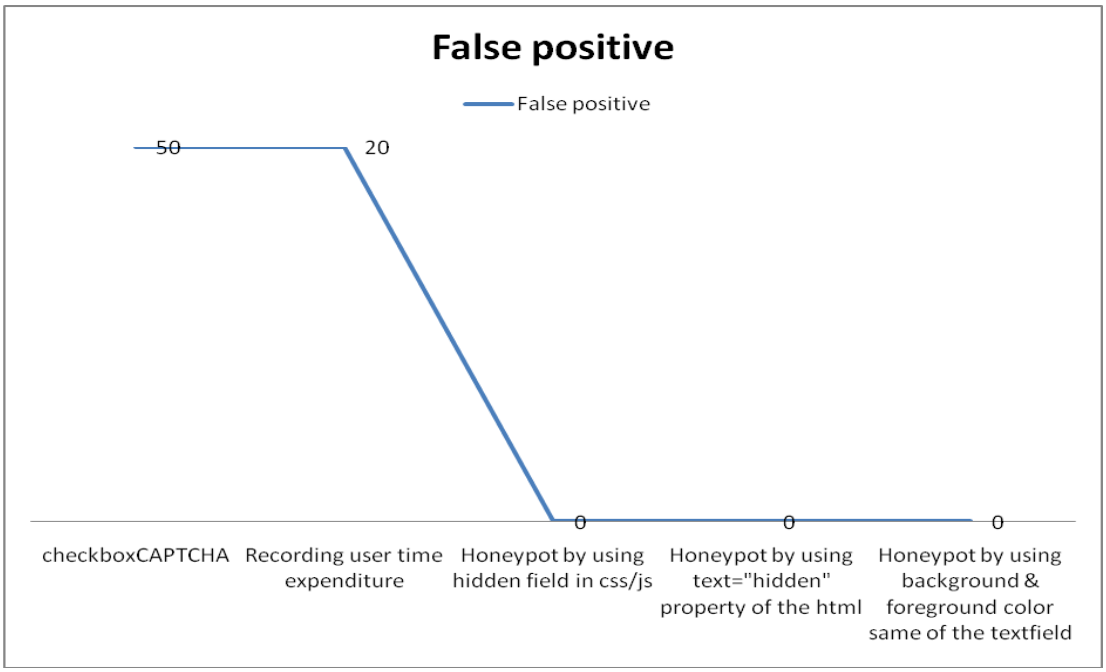


*Figure 16 Graph of SPAM filtering techniques vs false positive*

## VI. Conclusion and Future Work

It can be concluded that Honeypot technique is the most widely Spam prevention technique by which we can minimize the number of spammers and we can minimize false positive ratio too. We analyzed these methods based on false positive ratio to find the most efficient in terms of security and user friendliness and honeypot emerged as winner between CAPTCHA and Honeypot.

In our future work, we will think to implement Honeypot method on other than public and social networking CMS websites and enlarge the scope of use of Honeypot.

## References

[1]. Manish Saxena, P. M. Khan," Spamizer: An Approach to Handle Web Form Spam" 978-9-3805-4416-8©2015IEEE, 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)

[2]. Yi Zhou, Kai Chen, Li Song, Xiaokang Yang, Jianhua He," Feature Analysis of Spammers in Social Networks With Active Honeypots: A Case Study of Chinese MicrobloggingNetworks " 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 978-0-7695-4799-2© 2012 IEEE

[3]. Firkhan Ali Bin Hamid Ali, Farhana Bt. Karim ,"Development of CAPTCHA System Based on Puzzle" 2014 IEEE 2014 International Conference on Computer, Communication, and Control Technology (I4CT 2014), September 2 -4, 2014 - Langkawi, Kedah, Malaysia, 978-1-4799-4555-9©2014 IEEE

[4]. Hassan Ishfaq, WaseemIqbal and Waleed Bin Shahid,,Attaining Accessibility and Personalization with Socio-Captcha (SCAP), 978-1-4799-6369-0©20 15 IEEE, roceedingsof 2015 12th International Bhurban Conference on Applied Sciences & Technology (IBCAST) 307 Islamabad, Pakistan, 13th -17th January, 2015

[5]. Paul Heymann, Georgia Koutrika, and Hector Garcia-Molina,"Fighting Spam on Social Websites: A Survey of Approaches and Future Challenges", 89-7801© 2007 IEEE,Published by the IEEE Computer Society

[6]. Tao Men, Yan Sun, Deming Wang, Deming Wang, Mingrong Wang," A Novel Dynamic CAPTCHA Based On Inverted Colors ", 978-1-4799-2716-6©2013 IEEE

[7]. Misako Goto, Toru Shirato, RyuyaUda," Text-Based CAPTCHA Using Phonemic Restoration Effect and Similar Sounds", 2014 IEEE 38th Annual International Computers, Software and Applications Conference Workshops.

[8]. Hsin-Chang Yang, Chung-Hong Lee," Post-Level Spam Detection for Social Bookmarking Web Sites", 2011 International Conference on Advances in Social Networks Analysis and Mining, 978-0-7695-4375-8© 2011 IEEE

[9]. Shailaja Tingre, Debajyoti Mukhopadhyay," An Approach for Segmentation of Characters in CAPTCHA", ISBN:978-1-4799-8890-7©2015 IEEE

 

[10]. "Honeypot Technique: Fast, Easy Spam Prevention", February 1, 2014, https://solutionfactor.net/blog/2014/02/01/honeypot-technique-fast-easy-spam-prevention/
[11]. "Captchas vs. Spambots: Why the Slider Captcha Wins", by anthony on 04/21/11,http://uxmovement.com/forms/captchas-vs-spambots-why-the-slider-captcha-wins/
[12]. "In Search Of The Perfect CAPTCHA",By David Bushell,March 4th, 2011http://www.smashingmagazine.com/2011/03/in-search-of-the-perfect-captcha/
[13]. "5 alternatives to CAPTCHA that won't baffle or frustrate users", March 19, 2014,http://www.experienceux.co.uk/ux blog/2014/03/19/5-alternatives-to-captcha-that-wont-baffle-or-frustrate-users/
[14]. "Contact Form Honeypots", February 27, 2015,http://pageaffairs.com/notebook/contact-form-honeypots
[15]. "Introduction to Spam and Anti-spam",http://www.beansoftware.com/Tutorials-Articles-Guides/Anti-Spam-Introduction.aspx
[16]. "Think Your Site Needs CAPTCHA? Try These User-Friendly Alternatives.", HANNAH ALVAREZ | APRIL 9, 2014"https://www.usertesting.com/blog/2014/04/09/think-your-site-needs-captcha-try-these-user-friendly-alternatives/
[17]. "Data Mining Techniques",https://sites.google.com/site/assignmentssolved/mca/semester6/mc0088/14

 