International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 05 | May-2018 www.irjet.net p-ISSN: 2395-0072

Assessment of Privacy Policies using Machine Learning

Ritav Doshi[1], Aditya Ahale[2], Gaurav Gharti[3], Prakhar Pathrikar[4], Dr. P.S. Dhotre[5]

[1234](Student, Department of Computer Engineering, Sinhgad Institute of Technology and Science (Narhe), Maharashtra, India)
[5](Professor, Department of Computer Engineering, Sinhgad Institute of Technology and Science (Narhe), Maharashtra, India)

Abstract— *The increase in no of services provided on the internet demands the need for awareness about online privacy. A privacy policy document is often very long and very technical in its explanation of the terms to the user. This results in complete denial of the user to read these policies. This paper proposes a machine learning based approach to simplify understanding of these policies. On visiting the privacy policy webpage of a service provider, our system will automatically grade the policy and check for the satisfaction of individual classes of privacy. This checklist and score can then be used by the user to judge whether the policy is privacy abiding or not.*

***Key Words*: Internet, Privacy, Privacy Policies, Machine Learning, Classes of Privacy**

1. INTRODUCTION

According to OECD [1], the concept of privacy is defined as follows: "It is the status accorded to data which has been agreed upon between the person or organization furnishing the data and the organization receiving it and which describes the degree of protection which will be provided." A privacy policy does this job of communication between the organization furnishing and the organization obtaining data. All the organizations that deal with the user's personal information are obliged to follow privacy regulations, but the methods relating to check this are really less. Without the information regarding the data collection and its use, the user is unable to make informed decisions.

The main aim of privacy policy documents is to provide the user a control over the information he/she is sharing with the service providers. Most of the times the choice is limited to either furnish information or not use the service. The terms listed in privacy policies are in a free text full of technicalities and formal language which are often incomplete [2, 3]. This leads to the user choosing the easier way of unconditional acceptance.

When talking about the Indian context the Indian IT laws [5] have to be followed by the IT service providers. Most of the IT act laws are already covered in the OECD guidelines, but still shedding a light on these would help improve the privacy awareness in the Indian context.

There is a need for a system which helps the user to make informed decisions about their online privacy and also have control over these decisions. We propose a tool which automatically generates a grade for the privacy policies user encounters. This grade is based on the OECD defined eight classes of privacy [4]. These categories are collect, choice, cookies, access, purpose, security, share, and retention. The completeness in satisfying each category contributes to the overall grade of the policy. All the categories thoroughly covered in the privacy policy are also displayed in a tabular format. This helps the user to refer the policy back to check if his/her interests are being served properly or not. The machine learning algorithm classifies the policy as being closest to one among 52 policies in our researched dataset. This dataset has pre-assigned values for each privacy category. The privacy policy fetched from the website is pre-processed and shortened. On the basis of this, a Trust Score is calculated.

The paper is organized as follows:

The paper has the following contents: Section 2 contains the Literature Review which elaborates on the previous works in the field, Section 3 gives the method and details of the survey; Section 4 describes the detailed analysis of the survey, Section 5 tells about the Conclusion and Future Work.

2. RELATED WORK

There are mainly two kinds of researches done in this relatively new field. The first is, surveying the users, service providers, and different stakeholders to get a general perception of the understanding of the privacy policies. The second constitutes the solutions for these understanding problems. This can be in form of tools for judging the contents of a privacy policy.

A tool for analyzing the contents of an arbitrary privacy policy, completeness analyzer [7]. This approach uses machine learning techniques to comment on the completeness of a privacy policy. It uses a method based on pre-annotated paragraphs. The policy is parsed as paragraphs and then these paragraphs are annotated using classification techniques of machine learning. Different weights are assigned to categories which are then used to generate a score. The drawback of this method is that paragraph summarization is not a very accurate method of classification. Only 40 policies were used as a training set which was concentrated in few website categories only. The work in [7] also doesn't

specify the criteria of evaluation used for annotating the paragraphs in the privacy policy.

ToS; DR [6] like the modern slang acronym TL; DR (Too long, didn't read), is an online website which has graded major websites' privacy policies from very good (class A) to very bad (class E). It is a static project with a support for specific websites only.

[8] Explores different keywords and key extraction algorithms for the domain of online privacy policies.

The online privacy policy used for reference in this paper do not follow the IIA 2008.All the terms and conditions mentioned in the privacy policy are in accordance with the foreign rules. The keyword and key phases all differ from a foreign privacy policy to an Indian privacy policy. Also, a thorough understanding of the user's understanding ability is missing.

The paper can improvise the keywords and key phrases that are to be extracted as in the most of the Indian privacy policy and to make the extracted keywords and key phrases verify in accordance with the IIA 2008. Also to provide this evaluated info in such a form that is understandable to non- technical user too. These key phrases do not include keys from privacy policies mended in accordance with the IIA 2008.

All the above-mentioned literature and tools have tried to make the user aware of their online privacy. However, grading a website's privacy policy or giving an abstract overview of the policy doesn't cater to the complete understanding of the collection methods and the depth of access which the service provider demands. Some of the services mentioned are just displaying the rating or information about only some of the policies. Few of them are applicable and focused towards businesses only. Even if we mention the overview of the policy there is a need to tell the user about the meaning of the general terms of the policy.

So a few points that we thought needed more attention and which we covered in our tool are

- Expanding the domain.
- Covering most sites.
- Accuracy.
- Providing category details.

3. METHODOLOGY

3.1. Corpus

To generate accurate results using our Trust Score generator tool, the policies for training the algorithm have to be carefully chosen. We collected 52 policies from 20 different domains like commerce, education, entertainment, news etc. This widened the amount and diversity of words that were used to train the classifier. To reduce the computation time these policy documents were pre-processed beforehand. The corpus was used in 70-30 format, in which 70% of the processed policies were used to train the machine-learning algorithm and the remaining 30% were used for testing the classifier.

To understand the user mindset about online privacy, we also conducted a survey. 105 users participated in it, out of which 90% were computer engineering students. The questions in this survey were carefully modeled to generate critical responses from all the type of users. The frequently visited websites were given a slightly greater weight while designing the corpus.

3.2. Pre-Processing

On an average, the amount of words in a privacy policy is 2500. To increase the accuracy of our classifier and reduce the size of policies, the redundant and unimportant words need to be cut out of the policy text.

We used different Natural Language Processing techniques for this like lemmatization, stop-words removal, synonym removal, stemming.

After pre-processing, the actual running time of the algorithm decreased substantially.

3.3. Classification

The privacy policy given by the user has to be classified into one of the 52 categories to generate a score. The weka library provides a bunch of algorithms for classifying various kinds of data. We used the Naïve Bayes approach in our algorithm. This algorithm is advantageous when the classification has to be done in terms of words. This algorithm tries to classify the policy into one class while giving equal weightage to all the features. It tries to eliminate the features in which the policy doesn't classify.

3.4. Scoring Databases

In order to generate a Trust Score, the policies in the corpus needed to be graded beforehand. For this, 2 databases were created one the policy database and one the rating database.

The policy database contained the name of the vendor whose privacy policy was being studied and the corresponding classes of privacy that policy satisfied. The second database contained the name-value pairs of attribute and the corresponding grade it got. This was done for the case of all the privacy classes/categories.

4. PROPOSED ARCHITECTURE AND ALGORITHM:

The main aim of this tool is to help the user understand the privacy policy in a better way. For this, the tool focuses

on two result factors. One is the score, and other, the details about the presence of details about different privacy classes. The definition of all the categories was also displayed to help the user choose the classes which were important for him/her.

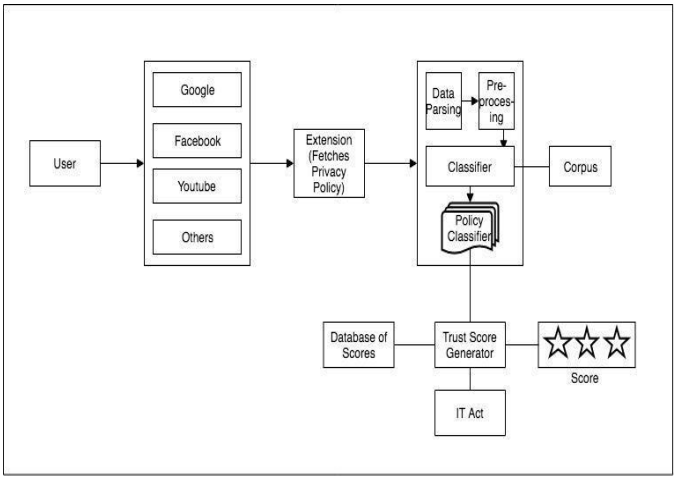

Fig.1 Trust Score generator tool- architecture

The components of our proposed architecture for the Trust Score generator tool are a browser extension, word pre-processor, classifier, corpus, database, score generator.

The user first opens the policy webpage of the service provider whose privacy policy he/she wishes to understand. On clicking the extension, it fetches the source code of the privacy policy webpage. This code is then separated from the HTML tags to generate the privacy policy text. The policy text is then cleaned with the help of different pre-processing techniques. This reduces the overhead on the algorithm.

This policy is given to the classifier. The Naïve Bayes classifier then classifies the policy using words as a feature. The algorithm labels this user is given policy as one of the policies in the corpus.

The categories satisfied by this labeling policy are then fetched from the database and the same values are assigned to the variables of the user given policy. The rating is then generated using the rating database. Most of the attributes for the privacy categories are in the form of three options viz. Yes, No and Not specified. According to the context of use in a category, these attributes are assigned a value from the set S= [0, 0.5, 1]. Adding values to all the privacy categories one final value is calculated. This value is then normalized to display on our Trust Score scale.

After all this server-side computation, the user gets to see the Trust Score of the policy and the classes satisfied by that policy. The user can hover the mouse on a particular category to get its definition.

4.1. ALGORITHM

4.1.1 Client-side algorithm

Step 1 Start Chrome.

Step 2 Visit any website's privacy policy page.

Step 3 On clicking the extension, fetch the page source of the webpage.

Step 4 Send website page source texts to the back-end server.

Step 5 Get the ratings from the server and display it on the panel.

Step 6 Show the checkboxes and put the tick or cross accordingly for classes of privacy satisfied

4.1.2 Server side algorithm to classify the policy text

Step 1 Fetch a privacy policy using Chrome API function.

Step 2 Pre-process the text and remove the HTML tags and irrelevant information.

Step 3 Apply stop word removal.

Step 4 Apply lemmatization and stemming.

Step 5 Pass this processed document to Naïve Bayes algorithm.

Step 6 Get the name of the policy (from the corpus) to which the user's policy bears similarity.

4.1.3 Server side algorithm for collecting class wise attributes from database

Step 1 Establish a connection with the MySQL database.

Step 2 Find the policy in the database in which the client-side policy is classified.

Step 3 Get the class wise attribute values and store them in the hash set.

4.1.4 Server side algorithm for generating the Trust Score

Step 1 Establish connection to the rating database in which contains attribute wise rating.

Step 2 Get the sore for each attribute

Step 3 Populate the scores

Step 4 Separate the classes with 0 value for attribute

Step 5 Normalize the score according to the display

4.2. RESULTS AND ANALYSIS

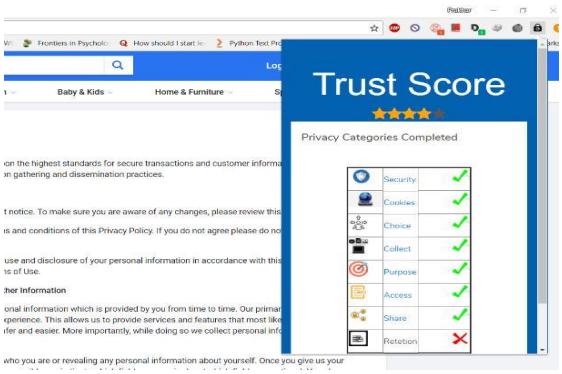

Fig.2. Trust Score generator tool- In working

Fig 2 shows the working of our extension. The above analysis is of the privacy policy of flipkart.com. According to our tool the policy is complete in all respects except the retention part.

On hovering over the cookie category, the information the privacy policy regarding cookies is shown.

The tool generates results in a matter of 3-4 seconds and works properly on 85% of sites we tested. The accuracy was >80% when we checked the same policies manually. For the policies in the training set accuracy was about 99%. This, also comments on the efficiency of our algorithm for fetching and separation of policy text from source code.

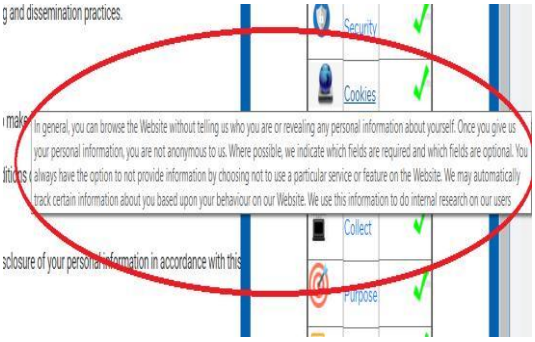

Fig.3. Details about the cookie section

5. CONCLUSION

The Trust Score serves as a medium of creating an understanding between the user and the service provider. It tries to put the user in control when the decisions regarding his/her privacy are concerned. Our tool works dynamically on most websites, but the structure of each website is different. This makes it difficult to scrape the policy text from this source code. All in all, Trust Score generator can serve as a great foundation for judging the privacy policy in a short time and take safe and unforced decisions about their online privacy.

6. ACKNOWLEDGMENTS

This work would not have been possible without the precise guidance and constant encouragement of Prof. Prashant S Dhotre.

REFERENCES

[1] http://oecdprivacy.org/

[2] A. Acquisti, I. Adjerid, L. Brandimarte. 2013. Gone in 15 Seconds: The Limits of Privacy Transparency and Control, EEE Security & Privacy, Vol.11, Iss: 4, pp.72, 74, July-Aug. 2013

[3] S. Zimmeck and S. M. Bellovin. Privee: An architecture for automatically analyzing web privacy policies. USENIX Security, 2014.

[4] OECD. Publishing. OECD guidelines on the protection of privacy and transborder flows of personal data. Organisation for Economic Co-operation and Development, 2002

[5] IndianITAct2008,http://www.eprocurement.gov.in/news/Act2008. pdf

[6] ToS;DR. Available: https://tosdr.org/

[7] E. Costante, Y. Sun, M. Petkovic, and J. den Hartog, A machine learning solution to assess privacy policy completeness, in Proceedings of the 2012 ACM workshop on Privacy in the electronic society - WPES 12, 2012, p. 91.

[8] Dhiren A. Audich ,Rozita Dara ,Blair Nonnecke, “Extracting keyword and key phrase from online privacy policies,” in Eleventh

[9] International Conference on Digital Information Management (ICDIM), Sept 2016.

[10] Alexa - Top Sites in India. [Online]. Available: http://www.alexa.com/topsites/countries;0/IN.