Thesis no: MSCS-2016-19

Text classification of Legitimate and Rogue online Privacy Policies

Manual Analysis and a Machine Learning Experimental Approach

Kaavya Rekanar

Faculty of Computing
Blekinge Institute of Technology
SE–371 79 Karlskrona, Sweden

This thesis is submitted to the Faculty of Computing at Blekinge Institute of Technology in partial fulfillment of the requirements for the degree of Master of Science in Computer Science. The thesis is equivalent to 20 weeks of full time studies.

Contact Information:
Author(s):
Kaavya Rekanar
E-mail: kare15@student.bth.se

University advisor: Martin Boldt
Title: Assistant Professor
Department of Computer Science and Engineering

Faculty of Computing	Internet	:	www.bth.se
Blekinge Institute of Technology	Phone	:	+46 455 38 50 00
SE–371 79 Karlskrona, Sweden	Fax	:	+46 455 38 50 57

Abstract

Context. A privacy policy is a statement made by any institution that informs clients about it's working practices. To avoid negative consequences, almost every institution (company) puts forth a policy that is supposedly written to protect a client's privacy. But, there may be some institutions that are intentionally set up to harm a client, like hosting spyware, lack of privacy, data laundering, etc. Usually, a client just agrees to the terms and conditions that are set up without even reading it, and that may be a scenario where a client is mislead. The policies that follow the rules set up Federal Trade Commission to write a policy have been considered as legitimate and those that have neglected those restrictions are considered as non-legitimate/rogue.
Objectives. This project deals on discriminating a privacy policy into a legitimate or rogue, depending on its content, using text classification mechanisms.
Methods. 100 legitimate and 69 rogue policies were collected and tested on 14 learning algorithms considering a baseline algorithm for comparison to check the performance by using a stratified 10-fold cross-validation test. The algorithms have been tested on two datasets, one which was collected just for the experiment and the other, was manually analyzed before sorting it out exactly into specific classes-legitimate and non-legitimate. The performance of both the datasets is tested using paired t-test.
Results. The 14 algorithms selected have given different classifier models based on their working, on both the datasets. The best performing algorithm in both the cases has been recorded as Naive Bayes Multinomial, hence chosen as the classifying model which could be developed into a tool.
Conclusions. A classification model has been developed which can give up 89.7% efficiency when classifying a policy based on its class, i.e., legitimate or rogue. The model's performance has been compared to baseline, in this study ZeroR is used to represent classification base on pure chance.

Keywords: Privacy policy, legitimate, rogue, text-classification, training, testing, learning algorithms

Acknowledgments

In this long page dedicated to a continual sequence of gratitude, I would like to state the harsh truth that I cannot actually name every person who has helped me throughout my exercise of thesis work in it's "sustained suffering" as the page would obviously not be enough. But, I would like to point out a few very special people who have been the backbone to the falling heap of my confidence.

The first person would definitely beyond the shadow of doubt be my adviser Dr Martin Boldt to whom I wish to express immense gratitude for the continuous support in my research work and for his patience(which could be more than a hundred tonnes, if measured I think), motivation, enthusiasm and great knowledge.

Also, I would like to mention my annoyingly sweet brother who was always very cautiously reminding me of the deadlines that needed to be achieved; and my father who taught me various things in my 22 years of living, especially time sense to be more precisely thankful about; and last, but not the least, my mother whom I had endless fights throughout the course only because I never ate on time.

I am taking this opportunity to let you, my very special and "unique" family know that your support is highly appreciated and valued, but most importantly loved by me as I think this would not be possible without your support, though you still wouldn't accept the fact that I am "insanely" amazing!!

Contents

List of Tables

Chapter 1

Introduction

A privacy policy is an affidavit that reveals the working of a certain firm that manoeuvres and operate a website that handles various amounts of client data, e.g., cookie files, etc. On one hand, these policies enable users to engage in transactions and interactions on the Internet, while on the other hand, abuses and leakage of this information could violate the privacy of their owners, sometimes leading to serious circumstances [6].

A privacy policy is always a matter of trust. Studies reveal that an estimation of 77% companies are putting up a privacy policy in recent times [7]. These policies differ greatly from one institution to another and they address many issues that are different than those the users care about [7].

Reading a privacy policy demands time, which is very valuable to any person. So, going through them is considered as an unnecessary task by most of them. There were days when people blindly trusted the companies which put up a privacy policy as a highly secure and a "good" company. But, some websites misuse this trust to their advantage by putting up the most commonly-known symbol of trust, which is the privacy policy, and misguide the clients.

A well-written(legitimate) privacy policy lets a customer know the rights to which they are entitled. A privacy policy lets a customer know beforehand about what they are getting into, before they agree to the Terms and Conditions of a website (or company). Most of the privacy policies are difficult to comprehend due to their long length or excessive use of legal jargon. But since, an institution is required to write what they will be doing in their privacy policies, a proper detection of the policy can actually help the client know if it would bring upon any danger. Yet, this information, as said earlier, is put using very well-crafted words, which is usually difficult to understand unless the person reading it is experienced or an expert in the same field.

Consequently, there is a need for a client to know the difference between a legitimate and a non-legitimate privacy policy in order to stay safe during the installation process itself, i.e., before the client agrees to the company's Terms and Conditions. If such methods help in detecting any harm that the usage of that particular software or any other product can potentially bring upon the client, it would assist in staying safe from non-legitimate issues before they are about to establish any connections.

Importance of Privacy Policies

Some industries like banks, medical professionals, etc are entailed by the US-EU Safe Harbor Framework to maintain a privacy policy. The regulations for policies are set up by the Federal Trade Commission, in order to design them, which if violated under any circumstances would lead to serious repercussions when examined.

If an industry does not belong to that regulated group, its marketing and customer regulation will be benefited if there is a privacy policy shown to the target audience. The prominent concept in today's world is that, any reliable website should maintain a minimum set of regulations which can be put up into a privacy policy.

Legal Risks with a Privacy Policy

If an industry belongs to a regulated category, it is mandatory that a privacy policy must follow all the setup regulations and must cover all the issues governing that industry. A failure in that task would lead to suspension of business license or fines at the least.

If the industry is not in the regulated category, there is no obligation to have a privacy policy, which would mean that there is no particular legal exposure to have a privacy policy. But, the fact that must be considered would be that a poor privacy policy would always be a liability to a company every time it tends to violate its own policy.

Need of a Privacy Policy

Assuming that the industry does not belong to regulated category, a privacy policy is not a necessity. A company cannot frame its policy based on it's competitors policy. But the company's policy makers cannot assume that the competitor has estimated the customer demand. They should survey the existing customer base to measure expectations. A privacy policy can only be adopted when it has been confirmed to augment customer retention and marketing.

In some industries, a privacy policy is not a necessity, but with other industries it is not a requirement as it is a necessity to stay atop in the market. As companies are progressively “going green” to help out with the environment, privacy policies are becoming a part of common business practice. Users of the websites are concerned about their privacy, and by applying a privacy policy a company is ensuring both safety and their privacy to the customers.

Chapter 2
Background

There are certain rules that have to be followed by a privacy policy. These rules have been setup after careful planning and negotiations by different people representing nations world-wide. All these rules have been combined together and have been set up as a framework, which came into existence with the name, US-EU Safe Harbor Framework.

The US-EU Safe Harbor Framework

In 2000, the European Commission and the US government created the US-EU Safe Harbor Framework [1]. In 2008, the Department of Commerce finalized negotiations with the Swiss Government and came up with a US-Swiss Framework, which went into action from February, 2009; and it also paralleled with the US-EU Framework [2].

To join the US- EU Safe Harbor, an organization should self-certify itself to the Commerce Department that it complies with the seven principles and its related requirements laid out by the Federal Trade Commission. The FTC makes sure to enforce the promises that the organizations make when they accredit that participate in the Safe Harbor Framework [5].
The seven principles in the US-EU Safe Harbor Framework[1] are:

1. **Notice**
   Individuals must clearly be notified by the organizations about the purpose for which information is collected and used [3]. The individuals must be provided with information on how they can contact the organization with any inquiries or complaints, the types of third parties to which it discloses the information and the possibilities and methods the organization offers to limit the use of individual's information and it's disclosure.

2. **Choice**
   Individuals must be given the opportunity to ***opt out*** if their personal

[1] This content has been referred from the US-EU Framework, 2014

information can be disclosed to a third party or used for any purpose incompatible to the original reason for which the said information has been collected by the organization, at the least, the individual should be authorized subsequently [3]. It is mandatory for affirmative or explicit issues, ***opt in*** choices.

3. **Onward Transfer (Transfer to Third Parties)**
   Any information to be disclosed by the organization should strictly adhere to the Notice and Choice principles. When an organization desires to give an individual's information to any third party, it can do so if it makes sure that the third party subscribes to the Safe Harbor Privacy Principles or is subject to the Directive or any other adequacy finding [5]. But if that is not the case with the third party, the organization can make a written agreement with the party stating that the third party will at least provide the same level of privacy protection by the relevant principles [4].

4. **Access**
   Individuals must always be accessible to their personal information that an organization holds and should be able to correct, amend and delete that information where it is not accurate, exceptional to the case where the expense of providing such an access would be disproportionate to the risk the individual's privacy in the case in question, or where the rights of any other person apart from the individual are being violated [2].

5. **Security**
   Organizations must take proper precautions such that the personal information of every individual involved shall be protected from loss, misuse and unauthorized access, disclosure, alteration, and destruction [2].

6. **Data Integrity**
   Personal information of an individual must be relevant for the purposes it is being used by the organization. An organization should take rational steps to ensure that the data is reliable for the intended use, accurate, complete and contemporary [4].

7. **Enforcement**
   To ensure consent with the Safe Harbor principles, there must be

   (a) readily available and affordable independent recourse mechanisms so that each individual's complaints and disputes can be investigated and resolved and damages awarded where the applicable law or private sector initiatives so provide [5];

   (b) procedures for verifying that the commitments companies make to adhere to the Safe Harbor principles have been implemented [5]; and

(c) obligations to remedy problems arising out of a failure to comply with the principles. Sanctions must be sufficiently rigorous to ensure compliance by the organization. Organizations that fail to provide annual self-certification letters will no longer appear in the list of participants [5].

Chapter 3

Research Methodology

This chapter is focused on understanding the outline of the research done for the project. This chapter gives an insight into the problem addressed, the aims and objectives, the research questions intended to be answered, the contribution by the researcher and the design that the research follows.

3.1 Problem Description and Motivation

There is a need to agree to the Terms and Conditions of a company's policies before you use any of it's products. While most of the people never read the policies, they may be walking straight into a trap or they may be protecting themselves because of the company's policy. But, it is a risk people willingly take just to save time reading them as most of them are from trustworthy institutions. And even if they are not, the client doesn't really care unless there is problem which they are personally subjected to.

To avoid such mishaps, it is always good to look through the privacy policies to stay on the safer side. In this technologically advanced, apparently "time-saving" world, we could always rely on a tool to do the job for us. This project aims on developing one such model which will be able to evaluate a certain privacy policy as legitimate or rogue and warn the client about what they are getting into before they "Agree" to the policy.

3.2 Aims and Objectives

The aim of this project is to find a working solution to discriminate rogue privacy policies from legitimate policies. This solution could be added to implement a client side browser add-on and later be developed as a complete software module for future work.

The objectives of this project are:

- Collect dataset of both legitimate and rogue policies.

- Manually investigate and categorize the instances in the dataset
- Evaluate suitable metrics to use.
- Evaluate suitable algorithms to include in comparison of the dataset.
- Investigating which machine learning platform/suite to use.
- Evaluate the feasibility of the experiment by testing it on the dataset previously collected.
- Investigate how to implement client side browser add-on.

3.3 Research Questions

A research question is a statement that identifies the phenomenon to be studied. The aim of our work is to answer the following research questions;

RQ1. How can legitimate privacy policies be distinguished from rogue privacy policies based on their content?

This question has been formulated to know the major characteristics of a policy which would help in categorising it in an efficient manner. The task is achieved using a manual analysis on the initially collected dataset, which clearly states the features that are mandatory in a privacy policy and also checks if they are present in that policy or not.

RQ2. To what extent can text classification algorithms distinguish the content in rogue privacy policies from legitimate?

The extent of differentiation has been tested using a few metrics on the whole dataset. To help improve the efficiency of classification, different tokenizers have also been used, and the configurations of the selected algorithms have been modified which is further discussed in the section 5.2.

RQ3. What differences between legitimate and rogue privacy policies can be learnt from the evaluated classification algorithms?

A total of fifteen algorithms have been chosen in this research work. But in order to know the differences between the two classes taken, three algorithms have been chosen, one from Bayes, Trees and Rules type respectively; and their working has been described clearly. The working of the algorithms depict the technique that has been used to classify the text given, which is a straight answer to the research question. This has been done in chapter 7.

RQ4. How can a browser add-on be constructed based on the text classification results in RQ2?

A description of how an add-on can be constructed using WebDriver in Firefox and Chrome has been written in the section Formulating a Browser add-on from the experimental results in chapter 7.

3.4 Contribution

After collecting the dataset, a list has been made of elements which will help distinguish a legitimate policy from a rogue one. To answer the first Research Question which focuses on finding the differences among the websites' policies; every policy collected is manually inspected to check if the organization genuinely follows all the principles stated in the International Safe Harbor Policy. The policies have been marked accordingly based on the matching of the respective criteria which has been discussed in chapter 5.

The algorithms to conduct the experiment have been selected based on their performance for classification when a certain dataset is given. The advantages of every algorithm's usage have been considered and analyzed before their selection. In every case, the main focus was on the better performing classifier algorithm for a set of configurations based on the collected dataset.

Two datasets have been made to test the efficiency of the project. The first dataset is the Source-based dataset whose collection has been described in detail in the subsection 5.1.1. As said above, to know the differences between legitimate and rogue privacy policies, a manual analysis has been done. The results of the manual analysis have led to the creation of a second dataset, which is described in subsection 5.1.2.

For a browser add-on to be constructed, a model had to be developed. This has been done using Weka 3.6.14, based on the best performing algorithm given the datasets. After the model has been saved, it has to be put into WebDriver, which allows the model to be developed into a browser add-on using different methods for both Firefox and Chrome. This formulation has been explained in chapter 7.

3.5 Research Design

This project mainly emphasizes on the differences between legitimate and rogue privacy policies. To answer RQ 1, it has been considered necessary to evaluate the policies manually and then find out the major similarities among them, so that they can be used to train the dataset for the experiment. Using the manual analysis another dataset has been created, the one which had policies sorted into legitimate and non-legitimate based on the analysis.

The experiment is carried out in two steps: Training and Testing.Training the dataset is done using both Weka Explorer; while testing is done for both the result sets using Stratified 10-fold Cross Validation method in Weka Experimenter.
To know the which dataset performs better in the experiment, we use a paired t-test to evaluate in a better manner.

Chapter 4

Related Work

There is no much research done in this area, as of now. But there are certain research works which have focused on classification of a certain dataset using text classification algorithms [19], and then a few more have also used machine learning approach to find out if there is any spyware in the data that has been taken [9]; but there has been no research on classifying the privacy policies. So, the related work is more focused on the importance of privacy policies in today's world, and many more details regarding the experiment that has been conducted to achieve the aim of this project, which can be found in the following sections.

4.1 Presumed Ineffectiveness of Privacy Policies

In 2003, the Online Privacy Protection Act has been enacted which requires website owners to post a statement of their policies regarding the collection and sharing of personal information in the California State Legislature [10] . Though the goal of this legislation was to create some transparency about the data collection practices and to help users make informed decisions, it does not regulate the substance of websites' practices [10]; they only need to disclose those practices [10].
Privacy policies have been rendered as ineffective due to several reasons;

First, due to the fact that Privacy Policies are difficult to read- Most of them are written in legal jargon that makes it difficult for an average person to read and understand, because of which most of them do not bother to read them [12].

The framing of privacy policies is such that, most of them lead customers to believe that their privacy is protected and concerned for [13]. A study found that "Users do not read privacy policies because they believe that they do not have to; to consumers, the mere presence of a privacy policy implies some level of often false privacy protection" [13].

Even if a consumer can understand the privacy policy, they are not interested in investing the amount of time required to read privacy policies [11]. A study has proved that it would take an average person about 200 hours a year to actually read the policy for every unique website visited in a year, not to mention the updated version of policies for sites visited on a repeated basis [11].

Even if they could understand the policy and make time to do so, there is not enough market differentiation for users to make informed choices [14]. Furthermore, many website policies are vague about what user information they are collecting and how it is going to be used. As they are all equally poor, the users have no viable alternatives [14]. This is a market failure.

Lastly, even if there was a market differentiation, it is not comprehensible that the users will protect themselves [12]. The potential danger are not salient, not to mention the fact that they are difficult to evaluate against the benefits of using a website [15][16][12].

Machine Learning for text classification is the foundation of document categorization, news filtering, document routing, and personalization [8]. In text domains, effective feature selection is essential to make the learning task efficient and more accurate [8].

In this study, the setting is for 2-class problems, i.e., legitimate and rogue. The features for distinction for both the classes have been the same. As discussed in Background, any policy that follows the rules set up by the Federal Trade Commission is known to be legitimate and the opposite falls into the other class, which is known as rogue.

4.2 Sentiment Analysis and Machine Learning

Natural Language Processing (NLP) is a vast area of Computer Science that deals with the interaction of Computers and Human Language [17] [18]. Most of the tasks involved in NLP are classification tasks, in which a classification function with the capability of defining the correlation between a certain 'feature' and a 'class' is attempted to be produced [17].

The classifier developed has to be trained with a training dataset, and then it can be used to actually classify documents. Determining its model parameters is called training [21]. If these are chosen correctly, the classifier will be able to predict the class probabilities of the actual document in a very precise manner with a similar accuracy as the training examples given [21].

After construction, the classifier, for instance, would be able to tell that a document containing the words "quantum" should be categorized as a Physics article, while documents with words "arbitrage" an "hedging" should be categorized as a Finance article. Another classifier would be able to tell that mails starting with "Dear Customer/Guest/Sir" (apart from your name) and containing words like "Great opportunity" or "a one-time offer" can be categorized as spam.

Another use of classifiers is Sentiment Analysis [19]. The purpose of this is to determine the subjective value of a text document, i.e., how positive or negative is the content of a document [20]. Unfortunately, the classifiers have not yet been able to succeed very well in this area. The major contribution for this failure is the subtleties of human language; sarcasm, irony, context interpretation, use of slang, cultural differences and the diverse ways in which an opinion can be expressed- subjective vs comparative, explicit vs implicit, etc [20].

The basic concepts involved in text classification are [42]:

- Tokenization
- Word Normalization
- Bag-of-words model

Tokenization

The process of chopping up sentences into smaller pieces (words or tokens) is called Tokenization [23]. This segmentation into tokens can be done with the help of decision trees, which contains information to correctly solve the issues that are possible to be encountered [23].
Some of the issues that may be encountered include;

- The choice for the delimiter will for most cases be a whitespace ("Pandas are cute" could be "Pandas", "are", "cute"), but when there is a whitespace between words ("Po is a Kung Fu Panda" could be "Po", "is", "a", "Kung", "Fu" "Panda") [24].
- Punctuation marks play a very important role in specifying the value of a certain context, for instance, '!' puts extra emphasis on the sentiment of the sentence, while '?' can mean uncertainity. Though many tokenizers are geared towards removing punctuations, for Sentiment Analysis a lot of information can be deduced from them [24].

Normalization

The reduction of every word to its base/stem form, by chopping off the affixes is called Word Normalization [19].
The issues that have to be noticed here are;

- Capital letters should be normalized to lowercase, unless it occurs in the middle of the sentence; this could indicate a personal noun, which may be important for classification [25].
- Apostrophe is a punctuation that makes classification tougher. "Pooh's hunny" should obviously be tokenized as "Pooh" and "hunny", but I'm, we're, they're should be made I am, we are, and they are. To make it more difficult, it can also be used as a quotation mark [25].

Bag-of-words

After the text has been segmented into sentences, each sentence has been segmented into words, the words have been tokenized and normalized, and then a simple bag-of-words model of the text can be implemented [24]. In this representation, only individual words would be taken into account, and give each word a specific subjectivity score [26]. This subjectivity score can be looked up in a sentiment lexicon [26]. If the total score is negative, the text will be classified as negative, and it it is positive, the text will be classified as positive [26].

4.3 Filtering of Features Selected

The overall procedure of selecting features for distinguishing is to score each potential feature according to a particular feature selection metric, and then take the best 'k' features. Scoring involves the occurrences of a feature in training legitimate and rogue class training set examples separately and then computing a function of these [9].

The first step of filtering is to eliminate words with rare occurrences, as they are most likely not going to aid in future classifications. Word Frequencies usually follow a Zipf distribution. Zipf is the frequency of each word's occurrence is proportional to $1/(rank^p)$, in which rank is its rank among words sorted by frequency and p is a fitting factor close to 1.0 [27]. Considering the fact that many distinct words do not occur more than once, eliminating such terms seems like a sensible filtering method as it promises more efficiency at the classification.

If eliminating rare occurrences of words based on a count from actual analysis of the whole dataset is done before hand, it just means that there is a leak to the test set at the training point itself. Hence, without investing more resources for cross-validation studies, as this particular practice is unavoidable, the fact that it does not particularly use class labels must be accepted.

Additionally, very common words like 'a', 'of', 'the', etc can also be eliminated due to fact that these terms, in no way actually help classification in a positive manner. This is where the “Stopwords” come handy. Stopwords are language and domain-specific. Hence, choosing a good filter of Stopwords can be a solution to overcome the above-discussed challenges [8]. Stemming or lemmatizing is a feature engineering option that merges various word forms such as plurals and verb conjugations into one distinct term [8].

A contributory feature engineering choice is the representation of the feature value. Usually, a Boolean indicator of the word occurrence in the document is sufficient. Other options available are finding the frequency of the word, the count normalized by the Inverse Document Frequency of the word. In some situations where the document lengths vary, a normalization of the counts is necessary (in this case, it is necessary as the length of policies vary).

The last choice in the selection of features is whether to rule out all the negative correlated features. Some studies are of the opinion that classifiers built from positive features are more robust in performance than those built from negative features. Besides, algorithms like Naïve Bayes Multinomial model has shown better results than the traditional Naïve Bayes model in many instances [9] [52].

4.4 Privacy Preferences Project

The Platform for Privacy Preferences Project (P3P) enables websites to express their privacy practices in an XML-based machine readable format than can be retrieved automatically and interpreted easily by user agents, by the means of P3P-enabled browsers [67]. As a first step, the website sends a machine-readable proposal of its privacy policies [67]. The proposal can be automatically parsed by a user agent and compared with the user’s privacy preferences, thus the users do not need to read the privacy policies of every website they visit [62]. Privacy Bird [63] and Privacy Finder [64] are examples of P3P user agents, able to compare P3P policies with user preferences. A limitation of the P3P policies is with users preferences. A limitation of the P3P is that it needs server-side adoption, which is not easily obtained: according to [65] only 20% of the websites amongst the E-Commerce Top 300 is P3P enabled [65]. Rogue actors defiantly won’t assist users by adding P3P policies to their sites.

Why is P3P useful?

As said earlier, P3P uses machine readable descriptions to report the collection and use of data [67]. Sites that implement such policies make their practises explicit and open them for scrutiny by the public [66]. With the help of smart interfaces, browsers are able to make the public understand these privacy practises (policies) [66].

In this manner, browsers will be able to develop a predictive behavior to block content like cookies, which would in turn, be giving a proper incentive for all the eCommerce sites to behave in a privacy friendly manner [66]; which would avoid the scattering of cookie-blocking behaviors based on individual heuristics imagined by the implementer of the blocking tool which will make the creation of stateful services on the Internet a pain because the state retrieval will be unpredictable [65] [66].

Chapter 5

Experiment

For the experiment, initially dataset has been collected, and then the two datasets have been trained and tested with 15 different algorithms using Weka 3.6.14. The performance of the classifier thus developed has been tested using a few metrics which has been reported in the following sections.

5.1 Dataset Collection

Two datasets have been collected to conduct the experiment, whose procedure is described in detail in the following subsections.

5.1.1 Dataset 1

The dataset used in the present study is a collection of 100 policies from assumed legitimate companies and 69 policies from companies with questionable reputation when it comes to respecting user privacy, i.e., rogue privacy policies. For the policies of legitimate companies, the Fortune 500 list, released in 2016 by the Fortune magazines has been taken as it ranks the top 500 largest revenue making companies for the respective fiscal year [38].

The list has been retrieved from the official website of the Fortune magazine. The privacy policies from the official websites of the top 100 companies names in the list has been collected as ASCII format documents.

Later, the collected policies have been extracted into unformatted text format and removing all the headers, footers and any additional information like hyperlinks other than the actual text content of the policy to use it for the experiment. Hundred (100) policies from assumed legitimate companies have been collected in this manner.

The rogue policies have been collected from different sources. There is a website named SpywareGuide, which list companies with questionable software on the Internet. A program has been written to crawl the site and automatically extract policies of websites with the given requirements. Fifty-four(54) policies have been retrieved in that manner, from an approximate of 2000 websites listed.

Another website named scumware.org lists resources with security problems, for researchers working in the field of security and malware detection. This website keeps an everyday record of threats reported and also provides statistical reports of the threats in a monthly fashion. The URLs listed in this website are mostly under working status and have been reported due to some serious security issues. Eight (8) policies have been retrieved from the list provided on this website. Some websites have already been blocked by the Service provider and permission has been denied as it can be harmful for the user to use them, therefore only eight out of 208 policies could be retrieved.

The third source of policies is a website named virustotal.com which happens to be an excellent source of websites even remotely containing harmful trojans or any other kind of viruses in them. This website routed to another one named, support.clean-mx.de which gives the latest list of viruses updates recorded worldwide from the usage of Internet every hour. The URL's obtained from this website led to the official sites which could provide the required rogue policies. Six policies have been collected in this procedure. Most of the routed websites were blocked as they were reported to be dangerous, hence the small number despite the large resources.

The last policy was collected from an organization's official website which was stated to have a very badly framed privacy policy by the Times report on Aug 6, 2015.

The initial plan was to gather a hundred legitimate and hundred rogue policies; but considering the difficulty in finding many rogue policies, a count of 69 policies has been decided. The details of the dataset's collection is given in Table 5.1.

Source	Class	Count	Tested	Collected Date
Forbes 500	Legitimate	100	100	23rd February,2016
Spyware Guide	Rogue	54	2000	13th September, 2013
Scumware.org	Rogue	8	208	1st March, 2016
Support.clean-mx.de	Rogue	6	150	3rd March, 2016
Times Report	Rogue	1	1	3rd March, 2016

Table 5.1: Table showing the records of collected Dataset

5.1.2 Dataset 2

The second dataset has been collected from the first dataset, after the manual analysis has been done. The dataset has been classified into categories of legitimate and non-legitimate based on their scores in the manual analysis. All the policies have been taken in the same pattern as they were collected for the first dataset.

The performance results of for each class are presented separately and then reviewed, which is followed by an analyzed and discussion for both the classes.

Every policy collected has been manually read and validated only if they strictly adhere to the rules put up by the Federal Trade Commission in February, 2016 [1], and then recorded accordingly. Adherence to the rules has been recorded with the value '1' and non-adherence has been recorded as '0'. If there is a partial adherence to the rules, it has been recorded with '0.5'.

A list of all the data that can be gathered by a firm, as stated in the respective privacy policies has been made with the title Gathered/Accessed Information. The list contains elements that are most frequently gathered by a firm. It has been done after the first analysis of the dataset manually.

Legitimate Policies

The results of Manual Inspection of 100 policies which have been classified as legitimate is shown in Table 5.2, Table 5.3, Table 5.4, Table 5.5, and Table 5.6. The table has been split into parts for better legibility to the reader. The mean achieved for the legitimate policies when the seven rules, i.e. Notice, Choice, Onward transfer, Access, Security, Data Integrity, and Enforcement is 6.896, which implies that most of the policies actually belong to the legitimate category.

A list of all the data that can be gathered by a firm, as stated in the respective privacy policies has been made with the title Gathered/Accessed Information. The list contains elements that are most frequently gathered by a firm. It has been done after the first analysis of the dataset manually. The mean achieved for the Gathered/ Accessed Information[1] is 11.177.

The total mean for both the categories is 18.0729; which is clearly a very good sign that most of policies have been divided into their respective classes, for the source-based dataset.

Manually-Analyzed dataset has been created with an intention to have a mean

[1] Information has been abbreviated as Info. in the table to fit in the contents.

of 20 so the classification of the policies into classes for training in the experiment is near perfect.

Rogue Policies

The result of Manual Inspection of 69 policies which have been classified as rogue is shown in Table 5.7, Table 5.8, and Table 5.9. The table has been split into parts for better legibility to the reader.

The mean achieved for the legitimate policies when the seven rules, i.e. Notice, Choice, Onward transfer, Access, Security, Data Integrity, and Enforcement is 3.245, which implies that most of the policies do not follow the rules, thus leading them to belong in the class rogue.

The mean achieved for the Gathered/ Accessed Information[1] is 10.200. The total mean for both the categories is 13.445455; which is clearly a very good sign that most of policies have been divided into their respective classes, for the source-based dataset.

From the results obtained in this manual analysis, the policies are now categorised again into their respective classes. Any policy which gives a sum above 6 in the rules section has been categorised as legitimate, and a policy with a sum below 6 would belong to the class rogue.

According to the US-EU Framework, a policy must strictly follow all the rules, but the rule Onward Transfer has been taken into consideration for the manual analysis; under the grounds that if onward transfer is not happening at all, it is good for client itself, as the client's information would not be passed on to anyone else, which is a good sign for any legitimate policy [5] [4]. Hence, a sum of 6 or 7 has been fixed for the classification of policies into their respective classes. In the end, the dataset 2 had 104 legitimate policies and 65 rogue policies after the manual analysis had been done.

5.2 Algorithm Selection and Configuration

This study is performed to examine if the privacy policy classification is attainable using supervised machine learning algorithms. In order to investigate on the said area, a diversified list of algorithms has been chosen, specifically trying to have at least one of a kind in the list, for instance, perceptron and kernel functions, instance-based learners, Bayesian learners, decision tree inducers, meta-learners, rule inducers, etc. Weka 3.6.14 has been used for algorithm implementations. The configurations have been kept default in most cases.

METRICS	ICBC	Bank of China	Berkshire Hathaway	JPMorgan Chase	Exxon Mobil	Petro China	General Electric	Wells Fargo	Toyota Motor	Apple	Royal Dutch Shell	Volkswagen Group	HSBC Holdings	Chevron	Walmart Stores	Samsung Electronics	Citigroup	Allianz	Verizon Communications
Notice	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Choice	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Onward transfer	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Access	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Security	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Data Integrity	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Enforcement	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Sum:	7	4	2	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
Gathered/Accessed Info																			
IP address	0	0	0	1	1	0	1	1	0	1	1	1	1	1	1	1	0	1	1
Contact Information	0	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
Ad customization	0	0	0	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1
Tracks interaction	0	0	0	1	1	0	1	1	0	1	1	1	1	1	1	1	0	1	1
Third party tracking	0	1	0	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1
User access	1	1	0	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1
Acquisition	0	0	0	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1
Data purchase	0	0	0	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1
Data retention	1	0	0	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1
Shares w/ affiliates	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Shares w/ contractors	0	0	0	1	1	0	1	1	0	1	1	0	1	1	1	1	1	1	1
Shares w/ third parties	0	0	0	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1
Sum:	2	3	0	12	12	1	12	12	2	12	12	11	12	12	12	12	10	12	12
Total sum	9	7	2	19	19	8	19	19	9	19	19	18	19	19	19	19	17	19	19

Table 5.2: Manual Analysis of the Source-Based Dataset, class Legitimate: Policies 1-20

METRICS	Bank of America	Sinopec	Microsoft	Daimler	AT&T	Gazprom	AXA Group	Nestle	Banco Santander	Ping An Insurance Group	Mitsubishi UFJ Financial	Johnson & Johnson	Total	Procter&Gamble	China Life Insurance	Bank of Communications	Google	Vodafone	BP	American International Group
Notice	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Choice	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Onward transfer	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Access	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Security	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Data Integrity	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Enforcement	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Sum:	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
Gathered/Accessed Info																				
IP address	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Contact information	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Ad customization	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Tracks interaction	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Third party tracking	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
User access	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Acquisition	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Data purchase	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Data retention	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Shares w/ affiliates	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Shares w/ contractors	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Shares w/ third parties	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Sum:	12	1	12	12	12	12	12	12	12	12	11	12	12	12	12	12	12	12	12	12
Total sum	19	8	19	19	19	19	19	19	19	19	18	19	19	19	19	19	19	19	19	19

Table 5.3: Manual Analysis of the Source-Based Dataset, class Legitimate: Policies 21-40

METRICS	Ita Unibanco Holding	IBM	BMW Group	Comcast	Commonwealth Bank	Pfizer	Goldman Sachs Group	BHP Billiton	MetLife	Novartis	Royal Bank of Canada	Siemens	Prudential	Anheuser-Busch InBev	Nippon Telegraph& Tel	Rosneft	Westpac Banking Group	Banco Bradesco	Softbank	Honda Motor
Notice	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Choice	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Onward transfer	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1
Access	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Security	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Data Integrity	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Enforcement	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Sum:	7	7	7	7	7	7	7	7	6	7	7	7	7	7	6	7	7	7	7	7
Gathered/Accessed Info																				
IP address	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1
Contact information	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Ad customization	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Tracks interaction	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Third party tracking	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1
User access	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Acquisition	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Data purchase	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1
Data retention	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1
Shares w/ affiliates	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1
Shares w/ contractors	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1
Shares w/ third parties	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1
Sum:	12	12	12	12	12	12	12	12	8	12	12	12	12	12	5	12	12	12	12	12
Total sum	19	19	19	19	19	19	19	19	14	19	19	19	19	19	11	19	19	19	19	19

Table 5.4: Manual Analysis of the Source-Based Dataset, class Legitimate: Policies 41-60

METRICS	General Motors	United Health Group	TD Bank Group	Intel	EDF	Ford Motor	Deutsche Telekom	BASF	Boeing	Industrial Bank	UBS	ANZ	Cisco Systems	Sumitomo Mitsui Financial	Zurich Insurance Group	China Minsheng Banking	Merck & Co.	Roche Holding	Citic Pacific	National Australia Bank
Notice	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Choice	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Onward transfer	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Access	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Security	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Data Integrity	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Enforcement	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Sum:	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
Gathered/Accessed Info																				
IP address	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Contact information	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Ad customization	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1
Tracks interaction	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Third party tracking	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
User access	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Acquisition	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Data purchase	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Data retention	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Shares w/ affiliates	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Shares w/ contractors	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Shares w/ third parties	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Sum:	12	12	12	12	12	12	12	12	12	11	12	12	12	12	12	12	12	12	12	12
Total sum	19	19	19	19	19	19	19	19	19	18	19	19	19	19	19	19	19	19	19	19

Table 5.5: Manual Analysis of the Source-Based Dataset, class Legitimate: Policies 61-80

METRICS	Shangai Pudong Development	Walt Disney	CVS Health	Telefonica	Oracle	Conoco Phillips	Sanofi	United Technologies	ING Group	Coca-Cola	China Citi Bank	Morgan Stanley	Hewlett-Packard	Nissan Motor	GDF SUEZ	PepsiCo	Lloyds Banking Group
Notice	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Choice	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Onward transfer	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Access	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Security	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Data Integrity	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Enforcement	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Sum:	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
Gathered/Accessed Info.																	
IP address	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Contact information	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Ad customization	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Tracks interaction	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Third party tracking	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
User access	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Acquisition	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Data purchase	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Data retention	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Shares w/ affiliates	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Shares w/ contractors	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Shares w/ third parties	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Sum:	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
Total sum	19	19	19	19	19	19	19	19	19	19	19	19	19	19	19	19	19

Table 5.6: Manual Analysis of the Source-Based Dataset, class Legitimate: Policies 81-100

METRICS	ABX Toolbar 1.0	Casino Rewards	Adware Deluxe	Click Alchemy	Ineb Helper	Commander Toolbar	Aimface	Coupon Bar	Kuaiso Toolbar	Dogpile Search Toolbar	One Step Search	Download Receiver	One Toolbar	Ebates Moe Money Maker	Opinion Bar	Advanced Email Monitoring	Platrium	Flow Go Bar	Precision Time
Notice	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Choice	0	0.5	0	0	0	0.5	0	0.5	0	0	0	0	1	1	1	0	1	0	0.5
Onward transfer	0	0	0	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1
Access	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
Security	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	1
Data Integrity	0	0	0	0	1	0	0	0	0	0	0	1	1	1	1	0	0	0	1
Enforcement	0	0	0	0	1	0	0	0	0	0	0	1	1	1	1	0	0	0	0
Sum:	1	1.5	1	1	4	2.5	2	2.5	2	2	2	4	6	7	7	3	3	2	4.5
Gathered/Accessed Info.																			
IP address	1	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Contact information	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
Ad customization	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Tracks interaction	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Third party tracking	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
User access	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
Acquisition	0	0	0	0	1	1	0	0	0	1	1	0	1	1	1	1	1	1	1
Data purchase	0	0	0	0	1	0	0	0	0	0	0	0	1	1	1	1	1	1	1
Data retention	0	0	1	0	1	0	0	0	0	0	0	0	1	1	1	1	1	1	1
Shares w/ affiliates	0	0	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1
Shares w/ contractors	0	0	0	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1
Shares w/ third parties	0	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1
Sum:	5	1	7	5	10	8	8	8	8	9	9	5	11	12	12	11	11	11	11
Total sum	6	2.5	8	6	14	10.5	10	10.5	10	11	11	9	17	19	19	14	14	13	15.5

Table 5.7: Manual Analysis of the Source-Based Dataset, class Rogue: Policies 1-20

METRICS	Adware Verticity	Protected Storage Passview	Red V	Brightest Flashlight	Search Words	Critical Stack Intel Marketplace	Smartpopups	HD Guru	Tone Locxx	Instagram	Top 20 results	Viewpoint Media Toolbar	LinkedIN	VS Toolbar	Lyft	WebMiner	Malware Domain Blocklist	Websearch	Snapchat	Wurld Media
Notice	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Choice	0	0	1	0	0	0	0	0	0	1	0	1	1	0	0.5	0.5	0	1	1	0
Onward transfer	1	1	1	1	0	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1
Access	0	0	0	0	0	0	1	0	0	1	0	0	1	0	1	1	0	0	1	1
Security	1	0	1	0	0	0	0	1	1	1	0	1	1	0	0	0	0	0	0	1
Data Integrity	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0
Enforcement	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0
Sum:	3	2	4	2	1	1	3	3	4	6	2	4	6	3	3.5	3.5	2	3	5	4
Gathered/Accessed Info.																				
IP address	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
Contact information	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Ad customization	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Tracks interaction	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Third party tracking	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
User access	0	0	0	0	0	0	1	0	0	1	0	0	1	0	1	1	0	0	1	1
Acquisition	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Data purchase	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Data retention	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1
Shares w/ affiliates	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Shares w/ contractors	1	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1
Shares w/ third parties	1	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1
Sum:	11	11	11	11	11	6	12	11	8	12	11	11	12	10	12	12	11	11	12	12
Total sum	14	13	15	13	12	7	15	14	12	18	13	15	18	13	15.5	15.5	13	14	17	16

Table 5.8: Manual Analysis of the Source-Based Dataset, class Rogue: Policies 21-40

METRICS	Summit Casing	Big Traffic Network	Twitpic	Ace Club Casino	Adware Safety	Atlas arbor	Hithopper	GotSmiley	Gator	Advertismen	FizzleWizzle Toolbar	eZulaDashConnect	File Freedom	DialerFactory	CowTrojan	DashBar
Notice	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Choice	0	0	1	0.5	0	1	0	0	0	0	0	1	0	1	1	0
Onward transfer	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Access	0	0	1	1	0	0	0	0	0	0	0	1	0	1	1	0
Security	0	0	0	1	0	0	1	0	0	0	0	1	0	1	1	0
Data Integrity	0	0	0	0	0	0	1	0	0	0	0	1	0	1	1	0
Enforcement	0	0	1	0	0	0	1	0	0	0	0	1	0	1	1	0
Sum:	1	2	5	4.5	2	3	5	2	2	2	2	7	2	7	7	2
Gathered/Accessed Info.																
IP address	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Contact information	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Ad customization	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Tracks interaction	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Third party tracking	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
User access	0	0	1	1	0	0	0	0	0	0	0	1	0	1	1	0
Acquisition	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Data purchase	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Data retention	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Shares w/ affiliates	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Shares w/ contractors	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Shares w/ third parties	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Sum:	11	11	12	12	11	11	11	11	11	11	11	12	11	12	12	11
Total sum	12	13	17	16.5	13	14	16	13	13	13	13	19	13	19	19	13

Table 5.9: Manual Analysis of the Source-Based Dataset, class Rogue: Policies 41-69

This study is not focused on finding an optimal configuration for every algorithm, i.e., the configuration for which the algorithm generates the best performing classifiers. If that was the case, it would demand an extensive parameter tuning. But this study is focused on investigating the possible differentiation of legitimate and rogue privacy policies. Therefore, any systematic parameter tuning has not been performed.
The selected algorithms and their configurations have been covered in a detailed manner in the section 5.4.

5.3 Classifier Performance Evaluation

Appropriate evaluation metrics have to be selected in order to measure classifier performance. The metrics selected have been discussed below.

Classification Accuracy

The Accuracy (the number of correct classifications/ total number of policies given) metric has been used conventionally. Nevertheless, several issues have been raised against the use of ACC as the only means of standard rating to measure performance [9]. But, when used with the other metrics in conjunction, ACC has been believed to be a useful metric.

Precision

While Accuracy is the degree of agreement between the experimental result and the true value; precision is the degree of agreement among a series of measurements of the same quantity, it is a measure of the reproducibility of results rather than their correctness [50].

Area under ROC curve (AUC)

AUC is an abbreviation for area under the curve. It is used in classification analysis in order to determine which of the used models predicts the classes best [50].
An example of its application are ROC curves. Here, the true positive rates are plotted against false positive rates. The closer AUC for a model comes to 1, the better it is. So models with higher AUCs are preferred over those with lower AUCs [50].

F-measure

F-measure(F1) is the harmonic mean of precision and recall [32]. F1 focuses majorly on the positive class and hence, negative features, even when inverted are devalued compared to the positive features [32].

5.3.1 Metrics

In order to focus on this classification study, six significant metrics are being considered, which are:

1. Classification Accuracy (ACC)
2. True Positives Rate (TPR)
3. False Positives Rate (FPR)
4. Precision
5. Area under the ROC curve (AUC)
6. F-Measure

True positive(TP) is a condition where "Legitimate" Privacy Policies are classified as "Legitimate". False positive(FP) is a condition where "Rogue" Privacy Policies are classified as "Legitimate". True negative(TN) is a is a condition where "Rogue" Privacy Policies are classified as "Rogue". False negatives(FN) is a condition where "Legitimate" Privacy Policies are classified as "Rogue". The equations show the formula used to calculate the true positive rate (5.1) and false positive rate of a classifier (5.2).

$$TruePositivesRate(TPR) = \frac{TP}{TP + FN} \tag{5.1}$$

$$FalsePositivesRate(FPR) = \frac{FP}{FP + TN} \tag{5.2}$$

5.3.2 Misclassification Costs

For the purpose of Policy Classification, we assume that the cost of misclassification is considerably different for both the classes, i.e., legitimate and rogue. For instance, classifying a legitimate policy as rogue is far worse than the opposite case and this is particularly true if the classification should be the basis for a decision support system that should aid the user in making the decision to install an application or to abort the installation [7].

If a legitimate policy is classified as a rogue, the user might opt for alternative software and go for it. On the other hand, if the opposite happens, i.e., if a rogue policy gets classified as legitimate, then the user may confidently install it under the wrong impression, thus becoming vulnerable to the repercussions that may be caused due to the wrong classification. This is actually worse than not having any classification done at all, since the user is actually believing in the legitimacy of the classification provided. Thus, this experiment is actually depicting a scenario where every classification is at a different cost, which means, different misclassification errors have asymmetric costs.

5.4 Experiment Procedure

Since there is a limited amount of data available for training and testing, the performance is estimated using the average of stratified 10-fold cross validation tests. Weka 3.6.14 has been used for all experiments and training the 15 learning algorithms on the two data sets [34].

The data collected is first converted into text files, removing all the headers, footers and unnecessary data (page numbers, logos, etc). As Weka accepts only .arff files, the dataset is converted into .arff and then used for conducting the experiment [34].

The filter StringToWordVector has been used on the dataset for preprocessing [35] [36]. The changes to the default settings that have been made in the process of filtering are as follows:

- Words has been set to 250, in order to pick 250 words per class, assuming that all the words are not important; this gives a little overlap.
- The Output Word Count has been changed from a default False to True. False tells a classifier "if" a word is in a document and True tells a classifier "how many" times a word is in a document [39].
- doNot OperatePerClassBasis has been set to True [39].
- Term Frequency (TF) and Inverse of the Document Frequency (IDF) value has been set to True.
  In information retrieval, TF and IDF are numerical statistic terms that are intended to show the importance of a word to a document in a dataset [40]. Hence, both of them have been set to True.
- The Stemmer chosen is Lovins Stemmer.
  Stemming has the capability to improve retrieval accuracy, but most of the stemmers are language specific [43]. Lovins Stemmer stems a word until it can no longer be reduced, using "bag of words" representation for the literature used in the dataset [41].
- The Tokenizer chosen is NGram Tokenizer.
  Character n-gram tokenization actually works in language-independent manner, but its use incurs a performance penalty [43]. If a word is considered as

a unit, alphabetic tokenizer works very efficiently. But testing the dataset using both NGram and Alphabetic tokenizers,on an average, it has been reviewed that algorithms are able to work efficiently when phrases are considered as a unit (i.e., when NGram Tokenizer is used).

- The StopwordsHandler used is Rainbow.
  StopwordsHandler is an interface for classes that support stopword handling. Rainbow has shown high performance in many researches [42], thus it has been used as the StopwordsHandler in this project.

The configurations for some algorithms used in the experiment have also been altered in order to yield better results, which are stated in Table 5.10.

Algorithm	Configuration
Baseline(ZeroR)	Default
Naive Bayes	Kernel estimator:False, Supervised discretization: False
NBMultinomial	Default
SMO	Kernel: Polynomial, C=1.0
Voted Perceptron	Exponent:1.0, Max kernel alterations: 10000
IBk	Number of Neighbors: 10, Distance weighing: false
KStar	Missing values treatment: average column entropy curves
AdaBoostM1	Classifier: Decision Stump
Bagging	Classifier: REP Tree
JRiP	Pruning: true, Number of optimizations: 2
PART	Binary splits:False, Pruning: True (confidence factor:0.25)
Decision Table	Default
Decision Stump	Default
J48	Pruning: Subtree raising, Pruning confidence factor: 0.25
RandomForest	Number of Trees:10

Table 5.10: Configuration changes in weka algorithms for the experiment

Cross-validation is a model validation technique to assess how the results of a statistical analysis will generalize to an independent data set [46]. It is usually used in settings where the goal is to predict, and one wants to estimate how accurately a predictive model will perform in practice. The goal of cross validation is to define a dataset to "test" the model in the training phase and give an overview on how the model will generalize in to an independent dataset in practice.

Thus, cross validation averages measures of prediction error to derive a more accurate estimate of model prediction performance. A Stratified 10-fold cross-validation test has been used in this project. Many tests have proved that Cross-Validation is actually better that repeated holdout [51]; it reduces the variance[2] of the estimate. 10 folds have been specifically used because, in the final analysis the entire dataset is going to be used for training; which means using 10-fold cross validation, 90% of the data is used for training, for 20-fold it would be 95% of the data. On the other hand, it has to be made sure that the evaluated data is a valid statistical sample; hence, it is not a really good idea to use a large number of folds for cross validation. Also, time constraint has to be taken into consideration where 20-fold would be taking twice the amount of time 10-fold took.

[2] Variance is a measurement of the spread between numbers in a data set. The variance measures how far each number in the set is from the mean.

Chapter 6

Results

The results of the experiment conducted on both the datasets- source-based and manually analyzed are presented in consecutive sections below.

As discussed earlier, ZeroR has been considered as the Baseline. The values which are marked ’v’ have better performance when compared to ZeroR; and those that are marked ’*’ are performing worse than ZeroR.

6.1 Experiment on the Source-Based Dataset

The results of the experiment on Dataset 1 are shown in Table 6.1.

Algorithm	ACC	TPR	FPR	Precision	AUC	F-Measure
ZeroR	0.600(0.024)	0.600(0.024)	0.600(0.024)	0.361(0.029)	0.500(0.000)	0.450(0.029)
Naïve Bayes	0.755(0.105) v	0.755(0.105) v	0.230(0.109) *	0.780(0.101) v	0.842(0.104) v	0.755(0.105) v
NB Multinomial	**0.839(0.081) v**	**0.839(0.081) v**	**0.179(0.095) ***	**0.851(0.078) v**	**0.897(0.075) v**	**0.837(0.082) v**
SMO	0.817(0.080) v	0.817(0.080) v	0.203(0.101) *	0.829(0.079) v	0.807(0.089) v	0.814(0.084) v
Voted Perceptron	0.823(0.088) v	0.823(0.088) v	0.207(0.110) *	0.832(0.097) v	0.864(0.087) v	0.818(0.096) v
IBK	0.765(0.100) v	0.765(0.100) v	0.239(0.113) *	0.782(0.098) v	0.782(0.115) v	0.763(0.102) v
KStar	0.792(0.086) v	0.792(0.086) v	0.228(0.101) *	0.803(0.087) v	0.858(0.085) v	0.789(0.087) v
AdaBoostM1	0.765(0.094) v	0.765(0.094) v	0.258(0.115) *	0.779(0.096) v	0.840(0.096) v	0.760(0.098) v
Bagging	0.775(0.101) v	0.775(0.101) v	0.273(0.128) *	0.791(0.102) v	0.848(0.101) v	0.765(0.109) v
JRip	0.719(0.115) v	0.719(0.115) v	0.304(0.129) *	0.735(0.124) v	0.718(0.126) v	0.710(0.122) v
PART	0.732(0.095) v	0.732(0.095) v	0.293(0.109) *	0.743(0.099) v	0.730(0.116) v	0.729(0.095) v
Decision Stump	0.687(0.114) v	0.687(0.114) v	0.285(0.127) *	0.727(0.117) v	0.701(0.118) v	0.685(0.116) v
DecisionTable	0.730(0.106) v	0.730(0.106) v	0.333(0.136) *	0.735(0.128) v	0.722(0.143) v	0.714(0.118) v
J48	0.724(0.105) v	0.724(0.105) v	0.302(0.128) *	0.737(0.111) v	0.726(0.121) v	0.719(0.108) v
Random Forest	0.819(0.081) v	0.819(0.081) v	0.258(0.121) *	0.851(0.069) v	0.882(0.082) v	0.802(0.098) v

Table 6.1: Results of the experiment on Dataset 1

All the chosen algorithms have given good performance when compared to the Baseline. The False Positive Rate should be lower than ZeroR in order to prove it’s efficiency. Other metrics have to be higher than that of ZeroR, which is exactly how the experiment worked out. Since there is class-imbalance problem in the dataset, i.e., the legitimate and rogue policies are not equal in number, considering the highest AUC value to be the base for measuring performance is the best method.

By looking at the results, it is clear that Naive Bayes Multinomial is the best model that can be developed to classify the privacy policies, with an AUC value of 0.897 maintaining a standard deviation of 0.075 for a 10-fold stratified cross-validation test. This shows that the classifier model developed from dataset 1 is showing an efficiency of 89.7% in classifying the text given.

6.2 Experiment on the Manually Analyzed Dataset

The results of the experiment on dataset 2 is shown in Table 6.2.

Algorithm	ACC	TPR	FPR	Precision	AUC	F-Measure
ZeroR	0.612(0.029)	0.612(0.029)	0.612(0.029)	0.375(0.036)	0.500(0.000)	0.465(0.036)
Naïve Bayes	0.829(0.085) v	0.829(0.085) v	0.207(0.104) *	0.836(0.087) v	0.881(0.095) v	0.825(0.088) v
NB Multinomial	**0.811(0.092) v**	**0.811(0.092) v**	**0.209(0.104) ***	**0.821(0.091) v**	**0.885(0.080) v**	**0.810(0.092) v**
SMO	0.774(0.085) v	0.774(0.085) v	0.255(0.093) *	0.783(0.085) v	0.759(0.087) v	0.771(0.087) v
Voted Perceptron	0.788(0.091) v	0.788(0.091) v	0.247(0.116) *	0.800(0.095) v	0.829(0.090) v	0.783(0.095) v
IBK	0.744(0.092) v	0.744(0.092) v	0.309(0.117) *	0.753(0.101) v	0.721(0.118) v	0.736(0.097) v
KStar	0.741(0.104) v	0.741(0.104) v	0.311(0.128) *	0.745(0.111) v	0791(0.128) v	0.733(0.108) v
AdaBoostM1	0.771(0.100) v	0.771(0.100) v	0.253(0.110) *	0.787(0.098) v	0.832(0.100) v	0.768(0.100) v
Bagging	0.756(0.100) v	0.756(0.100) v	0.321(0.127) *	0.774(0.106) v	0.823(0.105) v	0.741(0.107) v
JRip	0.724(0.099) v	0.724(0.099) v	0.300(0.108) *	0.742(0.098) v	0.719(0.103) v	0.720(0.100)
PART	0.735(0.096) v	0.735(0.096) v	0.296(0.114) *	0.747(0.098) v	0.694(0.136) v	0.731(0.096) v
Decision Stump	0.594(0.100) v	0.594(0.100) v	0.428(0.108) *	0.620(0.108) v	0.583(0.090) v	0.574(0.101) v
DecisionTable	0.719(0.095) v	0.719(0.095) v	0.366(0.124) *	0.728(0.116) v	0.722(0.138) v	0.701(0.105) v
J48	0.723(0.095) v	0.723(0.095) v	0.303(0.109) *	0.739(0.095) v	0.681(0.124) v	0.720(0.095) v
Random Forest	0.781(0.079) v	0.781(0.079) v	0.317(0.117) *	0.808(0.085)	0.858(0.097) v	0.759(0.095) v

Table 6.2: Results of the experiment on Dataset 2

By looking at the results, it is clear that Naive Bayes Multinomial is the best model that can be developed to classify the privacy policies, with an AUC value of 0.885 maintaining a standard deviation of 0.080 for a 10-fold stratified cross-validation test. This shows that the classifier model developed from dataset 2 is showing an efficiency of 88.5% in classifying the text given.

Chapter 7

Discussion

In this chapter all the research questions have been revisited and an explanation has been given on how each of them has been answered through the experiment.

RQ1: How can legitimate privacy policies be distinguished from rogue privacy policies based on their content?

According to the conditions put forth by the Federal Trade Commission in the US-EU Safe Harbor Privacy Policy, there are seven rules that every privacy policy must adhere to, which are: Notice, Choice, Onward Transfer, Access, Security, Data Integrity, and Enforcement [1]. When a policy follows all these rules, then it is considered as a legitimate; and if a policy is not following these rules, it has been considered to belong to the rogue class.

The process that is followed by a classifier model in order to classify the policies into their respective classes is shown in figure 7.1.

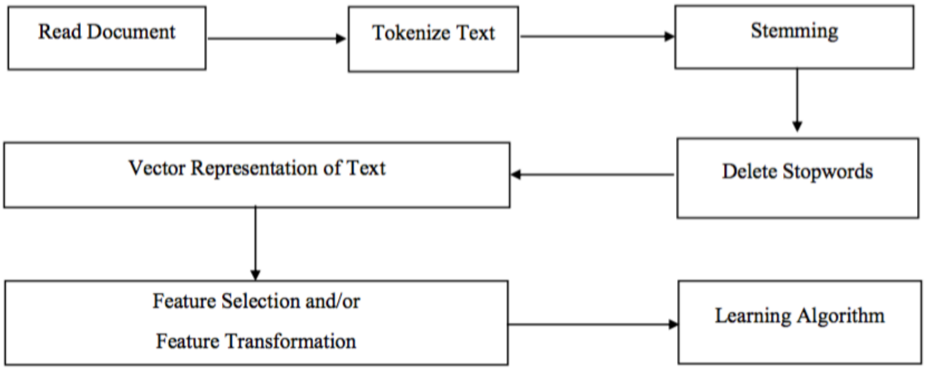


Figure 7.1: The working of a classifying model

There are two datasets that have been considered in this project: Source-based dataset, and Manually analyzed dataset.
The procedure for collecting these datasets has already been described in the

chapter Experiment. But, the technicality can be discussed here. The Source-based dataset (Dataset 1), as the name suggests, is actually a dataset that was collected initially from different sources. For legitimate, the privacy policies of the first 104 institutions listed in the Forbes 500 list have been taken, considering the fact that 4 of the top 100 did not have an online privacy policy. For the rogue class, the privacy policies have been collected from different sources (described in the section Dataset Collection), most of them with questionable content.

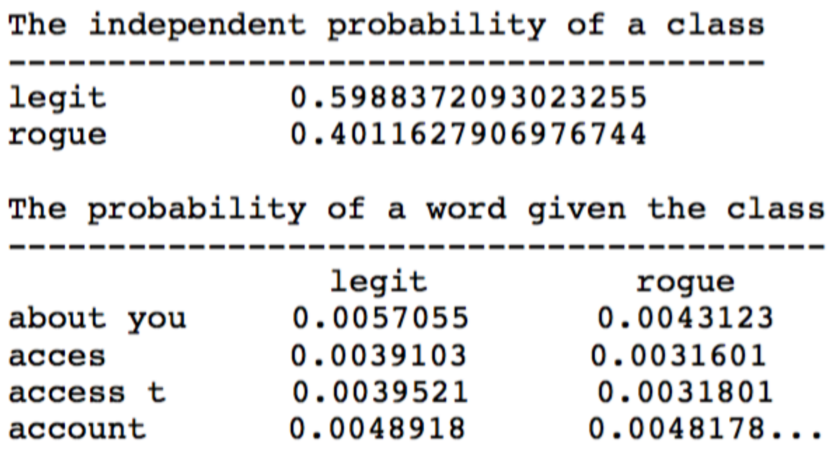
But, it was later in the Manual Analysis, it had been discovered that not all policies from either class follow all the 7 conditions, which lead to the formation of a second dataset from the results of Manual Analysis. The dataset collection has been explained in much more detail in subsection 5.1.2.

RQ2: To what extent can text classification algorithms distinguish the content in rogue privacy policies from legitimate?

According to the experiments conducted on both the datasets, it has been deduced that a text classification algorithm is able to distinguish a policy by 79.9%, which is the mean of AUC values for both the dataset. While Naive Bayes Multinomial has been the best classification model developed for both the datasets (0.885 and 0.897 respectively for a single hold-out), other algorithms have also performed very well when compared to the baseline (ZeroR).
Therefore, it is safe to say that text classification algorithms can distinguish the content in rogue privacy policies from legitimate to a large extent.

From the results it is clear that the Naive Bayes Multinomial model has been considered as the best model working. The reason is the working of the algorithm. It first calculates the independent probability of a class, and then the probability of every word given the class, as shown in figure 7.2 for the considered experiment [49].

The working of the algorithm is completely based on the basic rules of probability and permutations, which means that missing data will also be calculated as something missing only, and not be replaced by anything, like in Random Forest. So, on the whole, it can be considered that the Naive Bayes Multinomial model is the most suitable for development, as it handles the classification task with more efficiently than any other algorithm.



```
The independent probability of a class
--------------------------------------
legit          0.5988372093023255
rogue          0.4011627906976744

The probability of a word given the class
-----------------------------------------
                  legit          rogue
about you      0.0057055       0.0043123
acces          0.0039103       0.0031601
access t       0.0039521       0.0031801
account        0.0048918       0.0048178...
```

Figure 7.2: Naive Bayes Multinomial: Working of the model developed

RQ3: What differences between legitimate and rogue privacy policies can be learnt from the evaluated classification algorithms?

Three algorithms have been chosen to learn the differences between the evaluated classification algorithms. The algorithms have been chosen based on their type and performance in the conducted experiment; the chosen algorithms are- Naive Bayes Multinomial from Bayes, Random Forest from Trees and Decision Table from Rules, as these are the best performing algorithms in most cases of text classification [44].

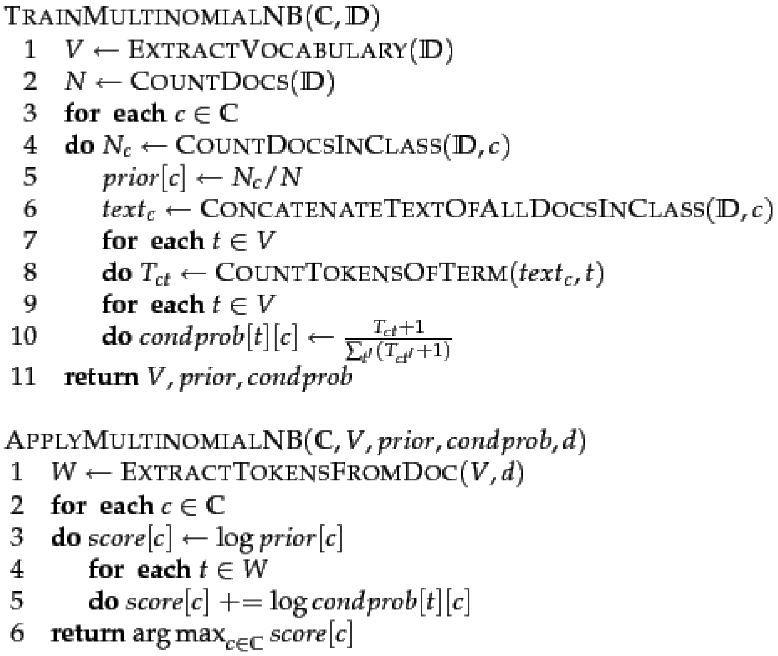
Naive Bayes Multinomial classifier

The Naive Bayes classifier is based on Bayes theorem with strong and naive independence assumptions. It is one of the most basic text classification techniques with various applications in email spam detection, personal email sorting, document categorization, language detection and sentiment detection. Naive Bayes performs well in many complex real world problems despite the naive design and oversimplified assumptions that the technique uses [55].

The efficieny of Naive Bayes classifier is because it is less computationally intensive (in both CPU and memory) and it requires a small amount of training data [55]. Morever, the training time with Naive Bayes is significantly smaller when compared to alternative methods [52].

Naive Bayes classifier is superior in terms of CPU and memory consumption, and in several cases its performance is very close to more complicated and slower techniques [52]. Usually, Multinomial Naive Bayes is used when the multiple occurrences of the words matter a lot in the classification problem [53]. The classifier follows the bayes rule for the classification process [54].

The training and testing algorithms for a Naive Bayes Multinomial classifier are presented in the figure 7.3.



```
TrainMultinomialNB($\mathbb{C}$, $\mathbb{D}$)
V ← ExtractVocabulary($\mathbb{D}$)
N ← CountDocs($\mathbb{D}$)
for each $c \in \mathbb{C}$
do $N_c$ ← CountDocsInClass($\mathbb{D}$, $c$)
    $prior[c] \leftarrow N_c / N$
    $text_c$ ← ConcatenateTextOfAllDocsInClass($\mathbb{D}$, $c$)
    for each $t \in V$
    do $T_{ct}$ ← CountTokensOfTerm($text_c$, $t$)
    for each $t \in V$
    do $condprob[t][c] \leftarrow \frac{T_{ct}+1}{\sum_{t'}(T_{ct'}+1)}$
return $V, prior, condprob$

ApplyMultinomialNB($\mathbb{C}$, $V$, $prior$, $condprob$, $d$)
W ← ExtractTokensFromDoc($V$, $d$)
for each $c \in \mathbb{C}$
do $score[c] \leftarrow \log prior[c]$
    for each $t \in W$
    do $score[c] \mathrel{+}= \log condprob[t][c]$
return $\arg\max_{c \in \mathbb{C}} score[c]$
```

Figure 7.3: Naive Bayes Multinomial: Training and Testing

Random Forest classifier

Random Forest is a versatile machine learning method that is capable of performing both regression and classification tasks [56]. It undertakes dimensional reduction methods, treats missing values, outlier values and other essential steps of data exploration, and does a fairly good job [57]. It is a type of ensemble learning method, where a group of weak models combine to form a powerful model.

In Random Forest, multiple trees are grown [56]. To classify a new object based on attributes, each tree gives a classification and the tree "votes" for that class [56]. The forest then chooses the classification that has the most votes over all the other trees in the forest and in case of regression, it takes the average of outputs by different trees [56].

Random forest is like a bootstrapping algorithm with Decision tree (CART) model [58]. Here, there are 169 observation in the complete population with 6 variables. Random forest tries to build multiple CART model with different sample and different initial variables [58]. For instance, it will take a random sample of 100 observation and 5 randomly chosen initial variables to build a CART model. It will repeat the process 10 times and then make a final prediction on each observation. The working of Random Forest classifier can be further understood from the figure 7.4.

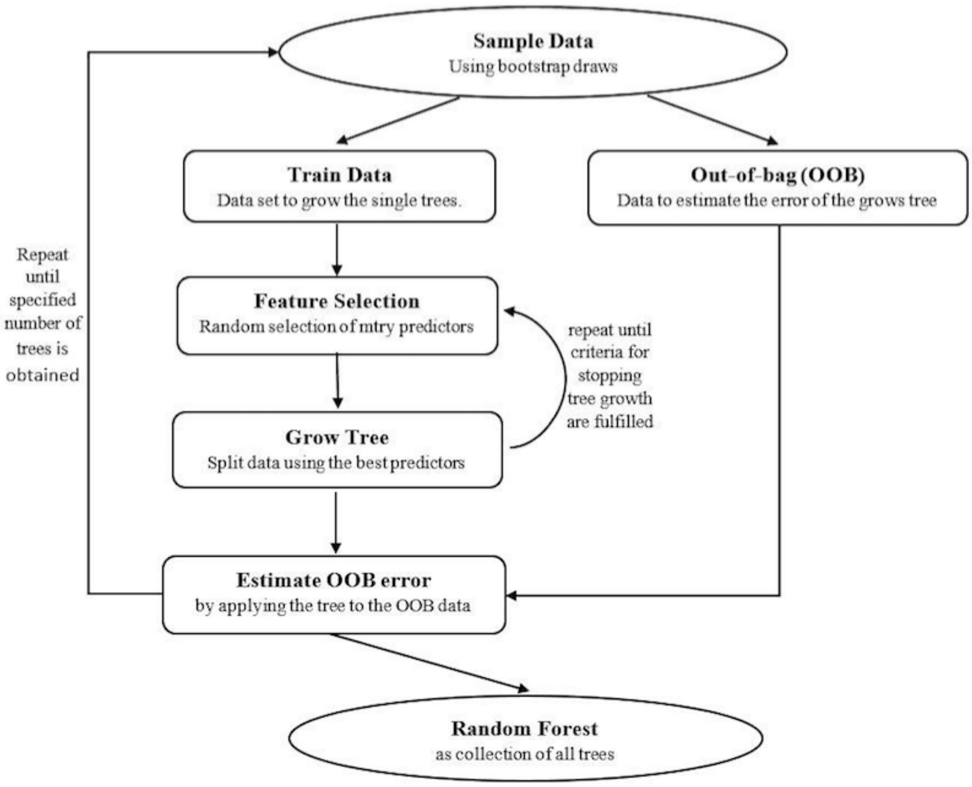

Sample Data
Using bootstrap draws
Train Data
Data set to grow the single trees.
Out-of-bag (OOB)
Data to estimate the error of the grows tree
Repeat until specified number of trees is obtained
Feature Selection
Random selection of mtry predictors
repeat until criteria for stopping tree growth are fulfilled
Grow Tree
Split data using the best predictors
Estimate OOB error
by applying the tree to the OOB data
Random Forest
as collection of all trees

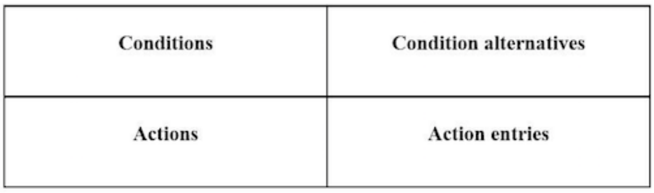
Figure 7.4: Random Forest: Training and Testing

Decision Table

A decision table is an excellent tool that can be used in both testing and requirements management [59]. Essentially, it is a structured exercise to formulate requirements when dealing with complex business rules [59]. Decision tables are used to model complicated logic [59]. They make it easy to look at all the possible conditions have been considered and when any conditions are missed, it is easy to see this [59].

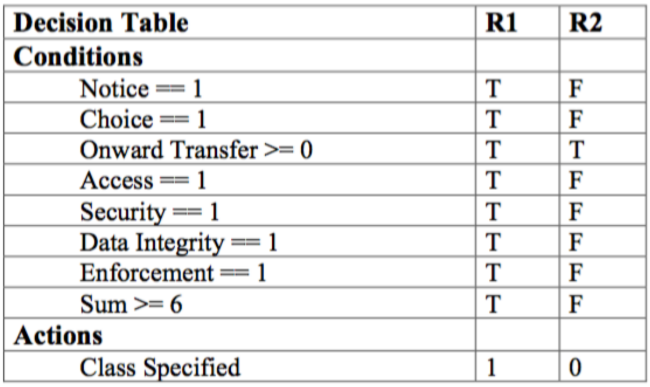
A decision table can be seen as four quadrants as shown in figure 7.5.
For better understanding of the concept, a decision table for Aimface policy has been depicted in figure 7.6.

In decision table, conditions are usually expressed as true (T) or false (F) [60].



Conditions	Condition alternatives
Actions	Action entries

Figure 7.5: Decision Table: Quadrants



Decision Table	**R1**	**R2**
Conditions		
Notice == 1	T	F
Choice == 1	T	F
Onward Transfer >= 0	T	T
Access == 1	T	F
Security == 1	T	F
Data Integrity == 1	T	F
Enforcement == 1	T	F
Sum >= 6	T	F
Actions		
Class Specified	1	0

Figure 7.6: Decision Table: Working

Each column in the table corresponds to a rule that describes the unique combination of circumstances that will result in actions which is shown in figure 7.6.

An advantage of using decision tables is that they make it possible to detect combinations of conditions that would otherwise not have been found and therefore not tested or developed [60]. The requirements become much clearer when a decision table is drawn, and it is often realized that some requirements are illogical, something that is hard to see when the requirements are only expressed in text [61].

A disadvantage of the technique is that a decision table is not equivalent to complete test cases containing step-by-step instructions of what to do in what order. When this level of detail is required, the decision table has to be further detailed into test cases [61].

Decision tables can be used in all situations where the outcome depends on the combinations of different choices, and that is usually very often [60]. In many systems there are tons of business rules where decision tables add a lot of value [60].

RQ4: How can a browser add-on be constructed based on the text classification results in RQ2?

Text classification involves two main steps:

- Representing the text database in order to enable learning, and training a classifier on it.
- Using the classifier to predict text labels of new, unseen documents.

The first step is a batch process, i.e., it can be done periodically. The second step is actually the moment in which an advantage is taken of the knowledge distilled by the learning process, and it is online in the sense that it is done by demand (when new documents arrive). This distinction is conceptual, that is, the modern text classifiers retrain on the added documents as soon as they get them, in order to keep or improve accuracy with time.

Adding ADD-ON in Firefox and Chrome

Assuming that the model to classify policies has been built using Save model after it has been trained, we can proceed to construct the add-on.

The model has to be saved on WebDriver to create an add-on on any of the browsers.

The steps to create an add-on in Firefox are listed below;

1: For this, a profile needs to be created.

2: Now, on the add-on's binary download page and download the add-on with .xpi extension and save it in Downloads.

3: Suppose taking path of .xpi file is c:/add-on, read the file location by using File class that is normally used for creation of files and directories, file searching, etc.

4: Call the addExtension() of FirefoxProfile class. This method will install add-on in new profile created with new Instance of Firefox.

5: Pass this profile in to new instance of FirefoxDriver.

As a whole, the code will look in Eclipse as shown in Listing 7.1.

Listing 7.1: Code to create an add-on in Firefox

```
FirefoxProfile firefoxprofile = new FirefoxProfile();
File addonpath = new File("path of .xpi file");
firefoxprofile.addExtension (addonpath);
WebDriver driver = new WebDriver (firefoxprofile);
```

The same steps cannot be followed for Chrome as we need to create instance of ChromeOption class. This class has many methods like addExtension(). This method is used to to install add-on in new instance of Chrome, setBinary(). This method is used to set the path to the Chrome executable, setArguments()- Adds additional command line arguments to be used when starting Chrome.
The steps to be followed in order to build an add-on in Chrome are,

1: Download the add-on in default location and read it using FileClass using:

```
File addonpath = new File("path of .crx file");
```

2: Now, create instance of ChromeOptions code using code snippet below.

```
ChromeOptions chrome =new ChromeOptions();
chrome.addExtensions(addonpath);
WebDriver driver = new ChromeDriver(options);
```

7.1 Usability of the model in a Practical Setting

The model (Naïve Bayes Multinomial) developed has given an accuracy of 0.87 for the first training, which means that it will improve if trained a number of times.

Using N-Gram tokenizer has helped in the classification to achieve a good True-positive rate, low false-positive rate, high precision, high AUC value and good F-measure.

Any policy needs to have the requirements discussed in Background Study in order to be classified as a legitimate policy. So, a policy maker must be aware of this and craft the policy in such a way to reach these levels, for it to be classified as legitimate. The model has been trained based on those rules only, and so, any policy that does not follow them will be classified as a rogue policy.

7.2 Validity Threats

7.2.1 Construct Validity

The experiment has been conducted using Weka 3.6.14 machine learning suite. So, if the experiment is conducted in any other way, the results may be different.

This training of the dataset has been done by using the most standard changes in the training phase for the dataset. If the same settings are used, and the dataset is trained using them, it would yield results that are closest to the results achieved from this experiment in any other suite as well, which does not change the validity of the experiment in a huge manner.

7.2.2 Internal Validity

Internal validity refers to the extent to which the researcher could claim that the independent variable caused the dependent variable [68]. The Internal Validity to this project is:

- The manual analysis may be biased, in some point of view.
  This threat has been mitigated by performing the manual analysis strictly according to new amendments made in the US-EU Safe Harbor Framework [2].

Chapter 8

Conclusions and Future Work

There are many websites available on the Internet that can create a privacy policy for an institution. But the policy created must adhere to all the rules set up by the Federal Trade Commission, so that the institution can be free from legal repercussions due to violence of the rules. And also, any client would be safe if a policy is bound by the conditions.

A model was developed in this project in order to check if a privacy policy is following all the rules in the US-EU Safe Harbor Framework. The model has been developed using Weka 3.6.14, a machine learning tool used to train and test a given dataset. The model that was developed was trained and test on 14 algorithms,apart from the baseline, out of which Naive Bayes Multinomial has been able to classify the text much more efficiently when compared to the other 13 algorithms. The Naive Bayes Multinomial classifier model gave an efficiency of 89.7% with an AUC value of 0.897 when tested using a paired t-test. The model has been evaluated using a stratified 10-fold cross validation test to know it's performance at every fold; this was done with an intention to develop the model with best capabilities in classifying content, as misclassification costs are very high.

The model developed in this project helps decide if the policy is actually following all the rules set up the Federal Trade Commission, as it has been built accordingly. This model is useful for both clients and an institution, to check if the policy is legitimate or not. The institution can check for the policy's legitimacy, and make changes in it's policy if it is classified as not good enough. The model could be useful for a client to check the policy's legitimacy to know what they are getting into by agreeing to the terms and conditions.

This project also gives a brief insight into how the model developed can be made into an add-on that can be installed in a browser, so that the end user would be able to access it whenever needed.

Future Work

A research can be done on the classifications performed by every algorithm in the classifier models that were trained and tested. Any policy that has been repeatedly classified into the wrong class can be found out from careful study of the models built by every algorithm, and the reason behind this behavior can be theoretically and experimentally researched. This could also effectively reduce the misclassification costs of the model developed, thus mitigating the risk to a minimum.

If more than one model could be incorporated in the tool developed for classification, then the misclassification costs could incredibly reduce, and that would make the developed tool almost 99% efficient. The models that could be clubbed are the best performing algorithms when they are trained and tested. The model developed could, for instance, be a combination of a Bayes, a rule-based and a tree-based algorithm, as in this research, these algorithms have yielded the best results of all the other algorithms that have been tested with both the datasets.

Developing a browser add-on from the model can be a good addition to the research done in this area. The add-on would make classification very easy for every person accessing the Internet as they can check the legitimacy of a policy for every website visited. It would make the end-user feel more secure about the website's usage, as business is always a matter of trust; and that trust could be built from developing a simple classification model that will help enhance a client's user-experience as an end result.

References

[1] “EU Commission and United States agree on new framework for transatlantic data flows: EU-US Privacy Shield.” http://europa.eu/rapid/press-release_IP-16-216_en.htm.

[2] “Reform of EU data protection rules - European Commission.” http://ec.europa.eu/justice/data-protection/reform/index_en.htm.

[3] “The EU-U.S. Privacy Shield - European Commission.” http://ec.europa.eu/justice/data-protection/international-transfers/eu-us-privacy-shield/index_en.htm.

[4] FTC (Federal Trade Commission), "FTC Policy Statement on Deception", 1983. http://www.ftc.gov/bcp/policystmt/ad-decept.htm

[5] FTC (Federal Trade Commission), "Privacy Initiatives".http://www.ftc.gov/privacy/

[6] M. C. Mont, S. Pearson, and P. Bramhall, “Towards accountable management of identity and privacy: sticky policies and enforceable tracing services,” in 14th International Workshop on Database and Expert Systems Applications, 2003. Proceedings, 2003, pp. 377–382.

[7] C. Jensen and C. Potts, “Privacy Policies As Decision-making Tools: An Evaluation of Online Privacy Notices,” in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, New York, NY, USA, 2004, pp. 471–478.

[8] G. Forman, “An Extensive Empirical Study of Feature Selection Metrics for Text Classification,” Journal of Machine Learning Research, vol. 3, no. Mar, pp. 1289–1305, 2003.

[9] N. Lavesson, M. Boldt, P. Davidsson, and A. Jacobsson, “Learning to detect spyware using end user license agreements,” Knowl Inf Syst, vol. 26, no. 2, pp. 285–307, Jan. 2010.

[10] A.D.Miyazaki and S. Krishnamurthy, “Internet Seals of Approval: Effects on Online Privacy Policies and Consumer Perceptions,” The Journal of Consumer Affairs, vol. 36, no. 1, pp. 28–49, 2002.

[11] McDonald, Aleecia; Cranor, Lorrie Faith, "The Cost of Reading Privacy Policies", CyLab, Carnegie Mellon University, 2008.

[12] Anton, Annie, "The Lack of Clarity in Financial Privacy Policies and the Need for Standardization", IEEE Security & Privacy, vol. 2, no. 2, 2004.

[13] Hoofnagle, Chris; King, Jennifer, "What Californians Understand About Privacy Online", Samuelson Law, Technology & Public Policy Clinic, 2008. http://www.law.berkeley.edu/clinics/samuelsonclinic/files/online_report_final.pdf

[14] Cranor, Lorrie Faith, et al., “2006 Privacy Policy Trends Report", CyLab Privacy Interest Group, 2007.

[15] Acquisti, Alessandro; Grossklags, Jens, "Privacy and Rationality, Privacy and Technologies of Identity", 2006. http://www.dtc.umn.edu/weis2004/acquisti.pdf

[16] Acquisti, Alessandro; Grossklags, Jens, "What Can Behavioral Economics Teach Us About Privacy","Digital Privacy: Theory, Technologies and Practices", 2007. http://www.heinz.cmu.edu/ acquisti/papers/Acquisti-Grossklags- Chapter-Etrics.pdf

[17] Manning, Christopher D., and Hinrich Schütze. "Foundations of statistical natural language processing", Vol. 999. Cambridge: MIT press, 1999.

[18] Murphy, Kevin P. Machine learning: a probabilistic perspective. MIT press, 2012.

[19] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002.

[20] Turney, Peter D. "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews." Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics, 2002.

[21] Ikonomakis, M., S. Kotsiantis, and V. Tampakas. "Text classification using machine learning techniques." WSEAS transactions on computers 4.8 (2005): 966-974.

[22] Rogati, Monica, and Yiming Yang. "High-performing feature selection for text classification." Proceedings of the eleventh international conference on Information and knowledge management. ACM, 2002.

[23] Mcnamee, Paul, and James Mayfield. "Character n-gram tokenization for European language text retrieval." Information retrieval 7.1-2 (2004): 73-97.

[24] Pak, Alexander, and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." LREc. Vol. 10. 2010.

[25] Kouloumpis, Efthymios, Theresa Wilson, and Johanna D. Moore. "Twitter sentiment analysis: The good the bad and the omg!." Icwsm 11 (2011): 538-541.

[26] Liu, Bing. "Sentiment analysis and opinion mining." Synthesis lectures on human language technologies 5.1 (2012): 1-167.

[27] K. W. D. Bock, "Advanced Database Marketing: Innovative Methodologies and Applications for Managing Customer Relationships", Routledge, 2016.

[28] R. L. Plackett, “Karl Pearson and the Chi-Squared Test,” International Statistical Review / Revue Internationale de Statistique, vol. 51, no. 1, pp. 59–72, 1983.

[29] C. Lee and G. G. Lee, “Information gain and divergence-based feature selection for machine learning-based text categorization,” Information Processing & Management, vol. 42, no. 1, pp. 155–165, Jan. 2006.

[30] T. W. Anderson, “A Modification of the Sequential Probability Ratio Test to Reduce the Sample Size,” The Annals of Mathematical Statistics, vol. 31, no. 1, pp. 165–197, 1960.

[31] J. L. Neto, A. D. Santos, C. A. A. Kaestner, N. Alexandre, D. Santos, C. A. A, K. Alex, A. A. Freitas, and C. Parana, Document Clustering and Text Summarization. 2000.

[32] D. M. Powers, “Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation,” Dec. 2011.

[33] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, “Short Text Classification in Twitter to Improve Information Filtering,” in Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA, 2010, pp. 841–842.

[34] Dimov, Rossen. "WEKA: Practical Machine Learning Tools and Techniques in Java."

[35] Han, Pu, Dong-Bo Wang, and Qing-Guo Zhao. "The research on Chinese document clustering based on WEKA." Machine Learning and Cybernetics (ICMLC), 2011 International Conference on. Vol. 4. IEEE, 2011.

[36] Neutatz, Felix, et al. "Evaluating Acoustic, Textual and Grammar Features for Alcohol Classification". http://suendermann.com/su/pdf/essv2016.pdf

[37] S. S. R. Mengle and N. Goharian, “Ambiguity measure feature-selection algorithm,” J. Am. Soc. Inf. Sci., vol. 60, no. 5, pp. 1037–1050, May 2009.

[38] C. Liu, K. P. Arnett, L. M. Capella, and R. C. Beatty, “Web sites of the Fortune 500 companies: Facing customers through home pages,” Information & Management, vol. 31, no. 6, pp. 335–345, Jan. 1997.

[39] Yang, Yiming, and Jan O. Pedersen. "A comparative study on feature selection in text categorization." ICML. Vol. 97. 1997.

[40] Zulkifeli, Wan, and Wan Rusila. Term frequency and inverse document frequency with position score and mean value for mining web content outliers. Diss. Universiti Putra Malaysia, 2013.

[41] Scott, Sam, and Stan Matwin. "Feature engineering for text classification." ICML. Vol. 99. 1999.

[42] M. Rogati and Y. Yang, “High-performing Feature Selection for Text Classification,” in Proceedings of the Eleventh International Conference on Information and Knowledge Management, New York, NY, USA, 2002, pp. 659–661.

[43] “Arabic Text Classification Using N-Gram Frequency Statistics A Comparative Study (PDF) - Semantic Scholar.” https://www.semanticscholar.org/paper/Arabic-Text-Classification-Using-N-Gram-Frequency-Khreisat/4cbd81db83c6f70a741fd1082ae3413133f8bd95/pdf.

[44] Ragas, Hein, and Cornelis HA Koster. "Four text classification algorithms compared on a Dutch corpus." Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1998.

[45] Eyheramendy, Susana, David D. Lewis, and David Madigan. "On the naive bayes model for text categorization." (2003).

[46] P. Refaeilzadeh, L. Tang, and H. Liu, “Cross-Validation,” in Encyclopedia of Database Systems, L. LIU and M. T. ÖZSU, Eds. Springer US, 2009, pp. 532–538.

[47] Chen, Chao, Andy Liaw, and Leo Breiman. "Using random forest to learn imbalanced data." University of California, Berkeley (2004): 1-12.

[48] Breiman, Leo. "Consistency for a simple model of random forests." (2004).

[49] Kibriya, Ashraf M., et al. "Multinomial naive bayes for text categorization revisited." Australasian Joint Conference on Artificial Intelligence. Springer Berlin Heidelberg, 2004.

[50] Davis, Jesse, and Mark Goadrich. "The relationship between Precision-Recall and ROC curves." Proceedings of the 23rd international conference on Machine learning. ACM, 2006.

[51] Kim, Ji-Hyun. "Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap." Computational Statistics & Data Analysis 53.11 (2009): 3735-3745.

[52] Huang, Jin, Jingjing Lu, and Charles X. Ling. "Comparing naive Bayes, decision trees, and SVM with AUC and accuracy." Data Mining, 2003. ICDM 2003. Third IEEE International Conference on. IEEE, 2003.

[53] Zhang, Haiyi, and Di Li. "Naive Bayes text classifier." Granular Computing, 2007. GRC 2007. IEEE International Conference on. IEEE, 2007.

[54] McCallum, Andrew, and Kamal Nigam. "A comparison of event models for naive bayes text classification." AAAI-98 workshop on learning for text categorization. Vol. 752. 1998.

[55] Rennie, Jason D., et al. "Tackling the poor assumptions of naive bayes text classifiers." ICML. Vol. 3. 2003.

[56] Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." R news 2.3 (2002): 18-22.

[57] Svetnik, Vladimir, et al. "Random forest: a classification and regression tool for compound classification and QSAR modeling." Journal of chemical information and computer sciences 43.6 (2003): 1947-1958.

[58] Elith, Jane, John R. Leathwick, and Trevor Hastie. "A working guide to boosted regression trees." Journal of Animal Ecology 77.4 (2008): 802-813.

[59] Sebastiani, Fabrizio. "Machine learning in automated text categorization." ACM computing surveys (CSUR) 34.1 (2002): 1-47.

[60] Lu, Hongjun, and Hongyan Liu. "Decision tables: Scalable classification exploring RDBMS capabilities." Very Large Data Bases: Proceedings, Cairo, Egypt, IEEE, New York, USA (2000).

[61] Gargantini, Irene. "An effective way to represent quadtrees." Communications of the ACM 25.12 (1982): 905-910.

[62] J. Reagle and L. Cranor, “The platform for privacy preferences: Communications of the ACM”, 42(2):48–55, 1999.

[63] L. Cranor and M. Arjula, “Use of a P3P user agent by early adopters”. In the 2002 ACM workshop on Privacy in the Electronic Society, pages 1–10, 2002.

[64] J. Tsai, S. Egelman, L. Cranor, and A. Acquisti, “The effect of online privacy information on purchasing behavior: An experimental study”, Information Systems Research, 21(June), 2010.

[65] P. Beatty, I. Reay, S. Dick, and J. Miller, “P3P Adoption on E-Commerce Web sites: A Survey and Analysis” IEEE Internet Computing, 11(2):65–71, Mar. 2007.

[66] Ashley, Paul, et al. "E-P3P privacy policies and privacy authorization." Proceedings of the 2002 ACM workshop on Privacy in the Electronic Society. ACM, 2002.

[67] E. Costante, Y. Sun, M. Petkovic, and J. den Hartog, “A Machine Learning Solution to Assess Privacy Policy Completeness: (Short Paper),” in Proceedings of the 2012 ACM Workshop on Privacy in the Electronic Society, New York, NY, USA, 2012, pp. 91–96.

[68] Randy L. Joyner, William A. Rouse, Allan A. Glatthorn, “Writing the Winning Thesis or Dissertation: A Step-by-Step Guide”, USA, Corwin Publication, 2013, Ch.9, p.p. 115-158.