Saint Petersburg State University

Graduate School of Management

Master in Business Analytics and Big Data

Development of a Website Classification Model for Quiet Media

Master's Thesis by the 2nd year students

Concentration – BM.5783.2020

Master in Business Analytics and Big Data (MiBA)

Gao Gan

Fu Songyuan

You Junwen

Research Advisor:

Vasilily Garshin

Olga Tushkanova

Saint Petersburg

2022

ЗАЯВЛЕНИЕ О САМОСТОЯТЕЛЬНОМ ХАРАКТЕРЕ ВЫПОЛНЕНИЯ ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ

Мы, _Гао Гань, Фу Сунюань, Ю Цзюньвэнь_, студент второго курса магистратуры направления «Менеджмент», заявляю, что в моей магистерской диссертации на тему «Разработка модели классификации сайтов для компании Quiet Media», представленной в службу обеспечения программ магистратуры для последующей передачи в государственную аттестационную комиссию для публичной защиты, не содержится элементов плагиата.

Все прямые заимствования из печатных и электронных источников, а также из защищенных ранее выпускных квалификационных работ, кандидатских и докторских диссертаций имеют соответствующие ссылки.

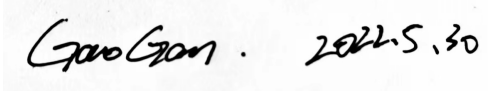
Мне известно содержание п. 9.7.1 Правил обучения по основным образовательным программам высшего и среднего профессионального образования в СПбГУ о том, что «ВКР выполняется индивидуально каждым студентом под руководством назначенного ему научного руководителя», и п. 51 Устава федерального государственного бюджетного образовательного учреждения высшего образования «Санкт-Петербургский государственный университет» о том, что «студент подлежит отчислению из Санкт-Петербургского университета за представление курсовой или выпускной квалификационной работы, выполненной другим лицом (лицами)».

STATEMENT ABOUT THE INDEPENDENT CHARACTER OF THE MASTER THESIS

We, Gao Gan, Fu Songyuan, You Junwen, (second) year master student, _MiBA_ program «Management», state that my master thesis on the topic «Development of a Website Classification Model for Quiet Media», which is presented to the Master Office to be submitted to the Official Defense Committee for the public defense, does not contain any elements of plagiarism.

All direct borrowings from printed and electronic sources, as well as from master theses, PhD and doctorate theses which were defended earlier, have appropriate references.

I am aware that according to paragraph 9.7.1. of Guidelines for instruction in major curriculum programs of higher and secondary professional education at St.Petersburg University «A master thesis must be completed by each of the degree candidates individually under the supervision of his or her advisor», and according to paragraph 51 of Charter of the Federal State Institution of Higher Education Saint-Petersburg State University «a student can be expelled from St.Petersburg University for submitting of the course or graduation qualification work developed by other person (persons)».



Gao Gan

Fu Songyuan

You Junwen

АННОТАЦИЯ

Авторы	Гао Гань Фу Сунюань Ю Цзюньвэнь
Название магистерской диссертации	Разработка модели классификации сайтов для компании Quiet Media
Факультет	Высшая Школа Менеджмента
Направление подготовки	Бизнес-аналитике и Больших данных
Год	2022
Научный руководитель	Тушканова Ольга Николаевна
Описание цели, задач и основных результатов	Цель исследования - построить высокопроизводительную модель классификации текстов для компании Quiet Media, чтобы автоматически дифференцировать сайты с неприемлемым контентом и оценить модель с помощью анализа рисков и анализа затрат, чтобы улучшить бизнес-продвижение рекламы в компании Quiet Media.
Ключевые слова	Телекоммуникационная промышленность, классификация веб-сайтов, TF-IDF, LSTM

ABSTRACT

Master students' names	Gao Gan Fu Songyuan You Junwen
Master thesis title	Development of a website classification model for Quiet Media
Title faculty	Graduate School of Management
Main field of study	Business Analytics and Big Data
Year	2022
Academic advisor’s name	Olga Nikolaevna Tushkanova
Description of the goals, tasks, and main results	The goal of the research is to build a high-performance text classification model for company Quiet Media to automatically differentiate sites with an inappropriate content and assess the model by risk analysis and cost analysis in order to improve business progress of advertising to company Quiet Media.
Keywords	Telecommunication industry, website classification, TF-IDF, LSTM

TABLE OF CONTENTS

Introduction

Since the 21st century, we have ushered in a big explosion of information, information overload has become the most severe problem of the Internet. According to the statistics of the number of real-time websites in “internet live stats,” as of April 2022, the number of websites on the Internet is as high as 1.9 billion, which is an astonishing number. Some invalid, blank, malicious, and inappropriate websites are also mixed into them. They not only increase the load of the Internet, but also bring great burdens to Internet users, it has become difficult to filter and identify relevant information nowadays.

Telecommunications carriers are organizations of telecommunications service providers that provide wireless voice and data communications to their subscribed mobile subscribers (Techoopedia, 2018). It plays the indispensable role in our information-driven lives. However, the rapid growth of information volume has become the biggest threat in the telecom industry. Slow growth in subscriber base leaves telcos challenged to improve business performance. It requires not only advanced technology but also prohibitive costs for telecommunication operators to maintain the stability of traffic to provide users with high quality service. The main advantage of the challenge in this situation is the volume of additional services provided to subscribers and optimization of operation activity.

Regarding to the Internet users, overloaded website information is undoubtedly a huge burden as well. The information indiscriminately been transmitted contains inappropriate, malicious websites and commercial advertisements, which are superfluous, useless, or even harmful to adolescence. Because of the advertisements banners full of the Internet, the network transmission speed is weakened, network environment is cluttered. On the other hand, although digital advertisers could run the commercial ads everywhere, they still must face the problem of the low efficiency in the business performance due to the fierce competition with others. They could not focus on the target customers but to afford the high expense on digital advertisements.

Therefore, the telecommunication developer industry came into being. Developer for telecommunication companies provides comprehensive, healthy, multifaceted services in order to profile subscribers in terms of collected data to precise target advertising, provide banner blockers service to reduce the advertisement and create greener Internet environment and help to filter spurious tracking requests and transform telco into a valuable advertisement market player.

Since the developer for telecommunication companies provides a variety of services for multifaceted clients, despite the fact of the necessity of the exist of this industry, there are few academic papers regarding analysis and research of it.

Our project was initiated by the company Quiet Media, which is one of the developers of telecommunication operators. Our research provides a model that will help company to differentiate sites with an inappropriate content and will improve business process of advertising. It relies on machine learning and deep learning techniques to build a binary classification model.

The main research goal of this study is to build a high-performance text classification model for company Quiet Media to automatically differentiate sites with an inappropriate content and to assess the performance of the model based on risk analysis from the business perspective view.

The project task of our group is to provide technical improvement for company Quiet Media in the process of providing services to telecommunication operators. Firstly, we thoroughly proceeded the research in the advertising market and introduced the operation mechanism of the company Quiet Media. Secondly, focus on the problem to be solved, we conducted the related literature research to figure out the optimal solution to our project. After the preparation work, based on the data set which was provided by our project sponsor, we started to get acquainted with the data set, by data collection and prepossessing to clean and integrate the data, then develop a classification model for website classification. Additionally, based on model evaluation and risk analysis from business perspective view to implement model deployment, to provide business recommendations for Quiet Media.

To achieve the goal, it was decided to initiate the project to tackle the following tasks:

- collect and preprocess data for further model development by applying data crawling, data cleaning, data construction, feature selection and data transformation procedures (Data preprocessing stage);
- develop model to classify the content of the websites by selecting optimal classifier using specified evaluation metrics (Model development stage);
- apply risk and cost analysis methods to assess developed model from the business perspective view to consider its risks and challenges in the commercial market (Business analysis stage).

In pursuing those tasks, various tools of data preprocessing and model development were applied, including JupyterHub environment with Python libraries (Scikit-Learn).

The research has practical as well as business value. In terms of practical value, we create the classification model by identify the contents of the website to differentiate the toxic websites from the useful ones. As for the business value, the research applied the business knowledge based on the risk analysis and cost analysis to assess the model performance in the business process of advertising.

Overall approach for structuring work related to model development was defined to Cross-Industry Standard Process for Data Science (CRISP-DM) methodology. In addition, several natural language processing and machine learning methods were used for modeling purposes.

This research relies on primary data on 3 files of distinguished URLs (Uniform Resource Locator), which are called blacklist-HTTP-only, whitelist-HTTP-only, and all-HTTP-only. They are separately stand for txt files containing URLs with pornography contents, URLs without pornography contents and many random URLs. Theoretical background regarding digital advertising, machine learning methods were supplied by the academic papers, articles and studies. The research relied on technical documentation for Python programming language, ML and Big Data libraries, such as Scikit-learn.

Structurally, the research is divided into four chapters. The first chapter provides an overview of the business-related background. It defines research the goal, tasks, objectives. The second chapter provides a detailed literature research of the methods we employed in the data cleaning, data preprocessing and model development procedures. The third chapter is focused on the process of establishing the classification model and evaluation. The last chapter provides the analysis of business performance by implement the model in the business process.

Data science is organized and employed in this research according to the CRISP-DM methodology principles outlined in the work of Chapman, et al in 2000 (see **Table 1** below).

Table 1. Thesis structure within CRISP-DM methodology

CRISP-DM Phase	CRISP-DM tasks with related thesis chapter
Business Understanding	Problem Statement 1.4 The problem of inappropriate site filtering using site classification Determine Projects Objectives and Goals 1.5 Project specification Produce Project Plan 1.6 Project Plan
Data Understanding	Describe Data Problem & Data Introduction 3.1.1 Primary dataset introduction 3.1.2 Data type selection 3.1.8 Data problem
Data Preparation	Clean & Integrate Data 3.1.3 Data crawling 3.1.4 Data cleaning and integration Select & Construct Data 3.1.5 Data construction 3.1.6 Feature engineering and data transformation
Modeling	Build & Assess Model 3.2.1 Data split 3.2.2 Algorithms for web site classification
Evaluation	Evaluate Results, Review Process 3.2.3 Model quality evaluation metrics 3.2.4 Text based website classification models evaluation
Business Analysis	4.1 Market analysis 4.2 Project management 4.3 Cost analysis 4.4 Risk analysis

	4.5 Business recommendations

The project was collectively performed by the authors of the paper - Gao Gan, You Junwen and Fu Songyuan. Each task was addressed in a collaborative way, there was no strict division of project work by the authors. **Table 2** below represents the main types of work and the share estimate of workload invested in them by the authors.

Table 2. Distribution of authors

Type of work	Student	Workload
Industry and marketing overview and problem statement	You Junwen	40%
	Gao Gan	30%
	Fu Songyuan	30%
Literature review	Fu Songyuan	40%
	You Junwen	30%
	Gao Gan	30%
Datasets description and exploration	Gao Gan	50%
	You Junwen	25%
	Fu Songyuan	25%
Data Preprocessing	Gao Gan	70%
	Fu Songyuan	30%
Data collecting	Gan Gan	70%
	You Junwen	30%
Model building	Gao Gan	70%
	Fu Songyuan	30%
Model evaluating and optimization	Gao Gan	70%
	You Junwen	30%
Business analysis	You Junwen	50%
	Fu Songyuan	50%

Chapter 1. Project statement

1.1 Review of online advertising market

In recent years, Internet has astonishingly developed, it has become the main source of information dissemination. In the meanwhile, online advertising rapidly emerged as a new form of online business. It has gradually dominated the whole advertising market and replaced the traditional media. The management mechanisms for online advertising are the same as those used in other traditional advertising channels (such as newspapers, radio or television), but are more creative in terms of providing targeted and personalized advertising (V. Yurovskiy, 2015).

After accounting for the COVID-19 pandemic impact on people's work and lifestyle, the massive growth in online traffic has created a large potential of the online advertising market. Generally, online advertising has become a marketing strategy employed by various organizations that involves using the Internet as an intermediary to obtain website traffic, deliver marketing information to the target customers. According to Statista, in 2020 online advertisements accounted for over 53% of the total advertising expenditure (Russian Association of Communication Agencies, 2021). Moreover, the Russian advertisers' spending on digital ads is predicted to increase from 3.9 billion U.S. dollars in 2020 to 5.4 billion U.S. dollars in 2024（PwC., 2021）.

According to Tchai Tavor, 2011, online advertising has its own unique advantages, it is more attractive, acceptable, and cost-effective than other media and full of interactivity. The low cost makes it possible to reach the target audience and find groups of consumers with similar interests. And interactivity allows users to express their reactions to ads through clicks. (Tchai Tavor, 2011)

Since the advent of online advertising, a variety of advanced technologies have been employed to meet the growing demand and solve problems faced by publishers and advertisers. Ad serving platforms are being introduced into the advertising ecosystem with the same intent, improving the entire media buying and selling process.

Advertisement server platform can be initially defined as a website server that stores and deliver the online advertisement to digital devices, such as websites and mobile applications. However, due to the rapid increasing demand for online advertising, basic functions of ad server platform no longer satisfied the users. Nowadays, modern ad server platforms could provide

advanced ad management solutions, including targeting, monitoring, evaluating, optimizing marketing campaign based on algorithms to profile target customers and etc.

The key players within the online advertising are advertisers, publishers and networks. Depends on the different parties in the online marketing process, Ad servers could be divided into 2 types, which are publishers' Ad server and advertisers' Ad server. In terms of publishers' Ad server, it allows website publishers to manage the ad slots and display sold ads. Regarding to advertisers' ad server, it helps advertisers to track campaigns, collect data and so on.

In addition, two parties, the technology systems also provide technical support in the online advertising process, which are known as DSP (Demand-side platform) and SSP (supply-side platform). DSP is a system that makes ad buyers to easily control the exchange of multiple ads and data exchange accounts by a single interface. While SSP is defined as a technology platform to enable web publishers to manage their advertising inventory, fill it with ads, and receive revenue (Internet Advertising Bureau,2021). They all consist of the complete interactions and play important roles within the online advertising progress. The interaction between key players in the online advertising market is summarized in the Figure 2 below.

Figure 1. Interactions within online advertising market

Source: [Author Research]

1.2 Quiet Media company and industry information

Quiet Media is a seller of advertising placed on the inventory of telecommunication operators founded in 2017 in Moscow, today is known as the largest provider of solutions for advertising placement and advertising traffic management in telecommunication operators' networks. The Russian telecommunication companies as the clients of company Quiet Media, such as MegaFon, Tele2, Rostelecom, Akado Telecom, Ecotelecom, which have subscribers using the services of operators cover the European part of Russia. From 2017 to 2021, Quiet Media has

established cooperation with Russian telecommunication companies and achieved the revenue more than 10 million dollars. (QuietMedia, n.d.)

The QIT platform developed by the company includes several segments, which are AdBlock.QIT, AdBanner.QIT, and Analytic. The functions unify the inventory of operators, targeting and visual appearance of advertising. Now the audience of the platform is about 96 000 000 unique users per month. By QIT platform online advertising, it has successfully completed commercial case through ad banner to attract online users to offline purchase on the operator's platform. (QuietMedia, n.d.)

1.3 QIT GLOBAL PAAS platform description

QIT GLOBAL PAAS is a platform developed by Quiet Media, combines the inventory of above-mentioned telecommunication operators and allows to contact subscribers in several Russian cities such as: Moscow, St. Petersburg, Murmansk and others. QIT Platform is developed with different functions and provide comprehensive services in the process of online advertising.

Within the AdBlock function providing a telecommunication company with a capability to control the advertiser's traffic reaching the subscriber. The telco became a targeted marketing communication channel with its customers, which developed high-precision targeting and visibility of advertising and transformed the telco into a valuable player in the advertising market. One of the functions called AdBanner.QIT, which is a technology providing telco with its own communication channel with the subscriber. Banners could be displayed in 2 formats, which are mobile formats and desktop formats ID QIT being a telecom integration platform that allows replacing fragmented identification tokens from various sources with a unified user ID. Partners are provided with a pixel that they install around them (on the website or in the application). Const. ID platform to identify the user and return the identifier to the partner. Adblock QIT enables telcos to separate useful traffic from harmful traffic, thereby reducing network load and allowing operators to capture a share of profits from active advertising streams. It is worth noting that our project was initiated based on this area to look for improvements in the business process of advertising.

In general, AdBlock as a network management tool, it helps telecommunication operators to clear the excessive spurious traffic, protect network environment and avoid network overload by filtering spurious tracking requests. The mechanism of AdBlock function has been demonstrated in the **Figure 3** below.

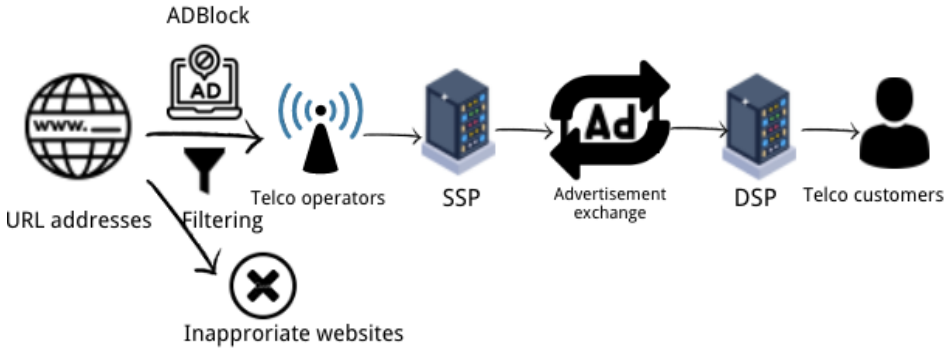


Figure 2. Demonstration of mechanism AdBlock function based on QIT platform

Source: [Author Research]

1.4 The problem of inappropriate site filtering using site classification

Based on the research of the online advertising industry and company business introduction, it was founded that it's crucial for company to provide customers with high quality and professional service within the dynamic marketing change. In addition, in order to satisfy the growing demand needs solid technical support behind it, Quiet Media has developed an advanced platform with a series of comprehensive service for online advertising, however, when publishing advertisements applying the mechanism of the platform, there is still room for improvement. The problem states below:

- Problem statement---Quiet Media company offers the indiscriminate URL addresses which include websites with inappropriate content for telco customers to advertise and inevitably cause the negative impact on the brand.

Our project was initiated to solve the above business problem by looking for improvements in the following research areas:

- Data exploration and visualization
- Construction of features that can be used for classification
- Exploring, building, and testing models for classifying
- Offer a tool that can be implemented into business process to manage all the tasks above

To expand the academic knowledge base of online advertising and solve the business problem of Quiet Media, this research will be focusing on addressing the following research goal:

- Research goal – to build a high-performance text classification model for company Quiet Media to automatically differentiate sites with an inappropriate content and assess the model by risk analysis and cost analysis to improve business progress of advertising to company Quiet Media.

To give the comprehensive understanding of the scale and accessibility of inappropriate websites on the Internet, our research from the beginning conducted research on it:

- Scale and popularity of the pornography websites-- Pornographic has been more accessible and popular than ever due to the Internet technology gives us the access to obtain massive amounts of information through Internet transmission, besides, the privacy of the network itself provide an ideal space for the spread of pornographic content. The computational neuroscientists Ogi Ogas, Ph.D., and Sai Gaddam, Ph.D., analyzed several billion recent Internet searches and found that of the 1 million most visited websites, 42 337 were sex-related, which is about 4 percent, the proportion that involved porn by the web searches from July 2009 to July 2010 is 13 percent, the porn accounts for around 10 percent of the material on the Internet. (Ogas & Gaddam, 2012). Overall, the problem of the huge accessibility to the pornography content on the Internet has been approved its seriousness.

The astonishing scale and popularity of pornography has become a biggest threat to the advertisers, as the online advertising developer company Quiet Media has the responsibility and solid technology to tackle this problem. Furthermore, the spread of pornography content caused the harmful consequence both business and society perspectives.

- Influence on society---According to the age distribution of pornography website in Russia in 2021, the largest share of pornhub.com visitors in Russia during 2021 were of age 18 to 24 years, represented by more than one third of the users. (Pornhub, December 2021). Besides, pornography has both primary and secondary effects on children and adults. It can make effects on adults' brain from the field of neuroscience and sexual aggression. (Byrin Romney, 2020) Moreover, adolescents who use pornography are more likely to have depressive symptoms and less emotional bonding to caregivers such as their parents. (Eric W.Owens, 2019).
- Influence on business and brand---From the business and online advertising industry points of view, on the one hand, appearance of advertisement on the pornography websites will cause the negative effects on brand image, brand attachment and brand

loyalty. Customers unwilling, even annoying to attach with their favorite brands in this way, which completely no benefits for marketing, but also make troubles. On the other hands, by transmitting these invalid ads to these pornography websites telecommunications operators must burden the extra traffic and remain the stability of the operation. It is undoubtedly uneconomical and unrealistic in the business.

In conclusion of this chapter, our research thoroughly analyzed the influence of the pornography websites' huge exposure to Internet users based on the research of scale and popularity to emphasize the necessity of this research on classification model study to filter them.

1.5 Project specifications

1.5.1 Gap analysis

Based on the problem of inappropriate site filtering using site classification and the research of the online advertising, several research gaps were identified:

- Insufficient introduction and definition of the industry of developer for telecommunication companies, such as company Quiet Media.
- Most website classification research lack of business financial value.

The meaning of our project not only filling the blank of the relative research but also of realistic meaning.

1.5.2 The goals and objectives of the project

Our main goal is to build a high-performance text classification model for company Quiet Media to detect the content of website and automatically classify whether it'as a Porn website or not. In addition to this, we are also expected to build the business model from the perspective of commerce, to analyze its practical meaning and the performance in business, to find the potential business value of the project.

To achieve the stated research goals, during the research process several tasks we planned to fulfill:

- Research task 1: Industry research and problem statement.
- Research task 2: Literature research of relative technical methods.
- Research task 3: Data collecting, prepossessing, model building and evaluation.
- Research task 4: Building business and financial model.

The expected result is the accomplishment of model building with the high performance and at the same time we evaluate the model with AUC-ROC curves to select the optimal one. The model should be accurate, fast to run, applicable into business and agile. Except from that, another expected result is a business model based on it, as our research is initiated by the real business, we assure our model satisfy the practical business use and predict the sustainability from the business perspective.

1.5.3 Overview of available IT resources

Table 3. IT Sources

Resource	Description	Application
JupyterHub	Web-based interactive computing platform. The notebook combines live code, equations, narrative text.	JupyterHub was used as a computational environment for machine learning and data mining. Tasks in Python programming language and its libraries.
Microsoft Excel	Spreadsheet program	Microsoft Excel was used to establish financial model.
GSOM server	Intel Xeon 6348H 24-core 48-line 128G DDR4 RAM 200T SSD ROM With Linux	GSOM server was used as a powerful and reliable server to perform computation in the process of data science research

Source: [Author Research]

1.6 Project plan

1.6.1 Project timeline

Our project consists of 4 chapters, the Gantt chart below specifies our project tasks by timelines (see **Figure 3** below). In the start of project, we also set a time plan for each work (see **Table 4** below). In the task table, green color means finished, yellow means in process and gray means canceled by some reasons.

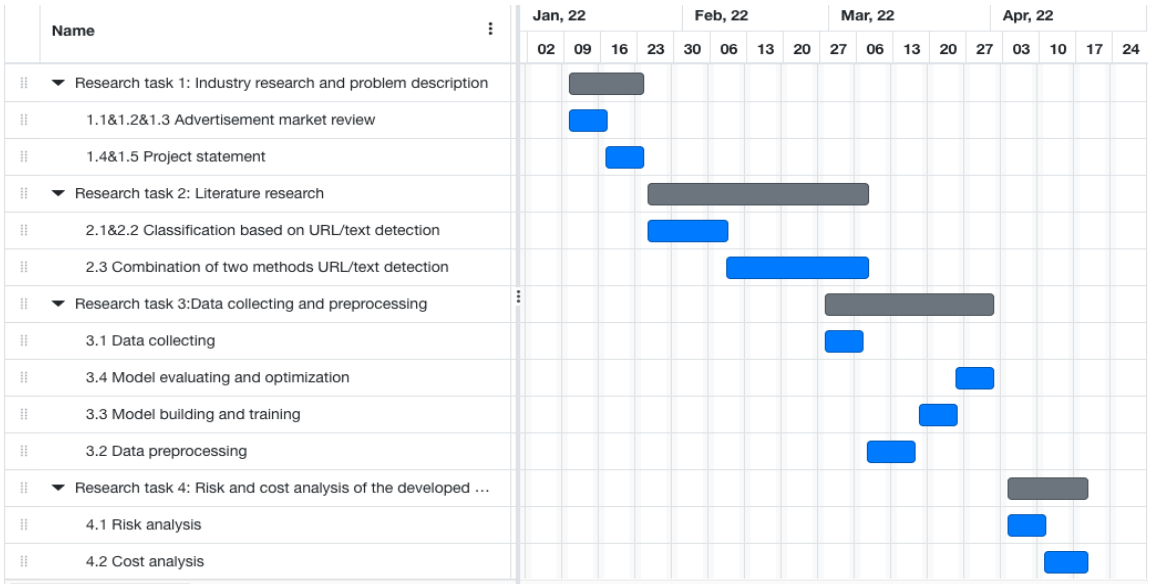


Figure 3. Gantt Chart of project

Source: [Author Research]

Table 4. Project task plan with running status

Week	Date from	Date Until	Learning Task	Main Task on Project	Other work	Team Meeting
1	10-Jan	16-Jan	Literature research	Prepare websites for training dataset		-
2	17-Jan	23-Jan	Python website crawler	Crawling text data of websites	Plan research	-
3	24-Jan	30-Jan	Beautifulsoup library for HTML text cleaning			27-Jan
4	31-Jan	6-Feb	Text preprocess methods for feature extraction	Feature extraction: clean text data and data preprocess	Draft of plan	3-Feb
5	7-Feb	13-Feb			MT Plan hand-in	10-Feb
6	14-Feb	20-Feb			Summary of literature and knowledge	17-Feb
7	21-Feb	27-Feb	ML model algorithm and accuracy test	Feature selection and adjust parameter		24-Feb
8	28-Feb	6-Mar		Accuracy test design		3-Mar
9	7-Mar	13-Mar		Training models and accuracy comparation		10-Mar
10	14-Mar	20-Mar				17-Mar
11	21-Mar	27-Mar	Literature research	Optimize and 1st edition output	Presentation for current results	24-Mar
12	28-Mar	3-Apr	Estimate model	Summary and prepare for presentation		31-Mar
13	4-Apr	10-Apr	DL network	Optimize and second edition output		7-Apr
14	11-Apr	17-Apr	Business analysis framework	Risk analysis and Cost analysis	Draft of MT	14-Apr
15	18-Apr	24-Apr	Financial Mode	Market analysis	Styling whole MT	21-Apr
16	25-Apr	1-May			Presentation for mock test	28-Apr
17	2-May	8-May	Master Thesis writting skill	Optimize and final edition output		5-May
18	9-May	15-May			Prepare for final Presentation Stroryline	12-May
19	16-May	22-May				19-May

Source: [Author Research]

1.6.2 Project team

The table below shows the main participants of the research project and their information.

Table 5. Project team statement

Name	Role	E-mail
Gao Gan	**Researcher**	st084310@gsom.spbu.ru
Fu Songyuan	**Researcher**	st091465@gsom.spbu.ru
You Junwen	**Researcher**	st084081@gsom.spbu.ru
Vasiliy Garshin	**IT supervisor**	vgarshin@gsom.spbu.ru
Olga Nikolaevna	**Academic advisor**	tushkanova@gsom.spbu.ru
Anton Ostroumov	**Sponsor**	

Source: [Author Research]

Chapter 2. Analysis of the current state of the market and technologies for inappropriate site filtering

2.1 Scientific approaches to inappropriate site classification

Since the emergence of the Internet in the late 20th century, web classification has been a topic of great interest and need, and many large Internet companies have attached great importance to this project, while scholars and researchers have been improving and optimizing the various processes of classification, including web information collection, data cleaning, text information processing, lexical and semantic extraction, selection and optimization of classification algorithms, and have made great contributions.

Through our reading and collection of literature, we have obtained a large number of methods to use and choose from, and we will subsequently combine the advantages and disadvantages of each method to find the ones that are most suitable for our project.

In this chapter we will discuss the literature research in various parts of the classification problem, including classification via URL, text and image. In addition, we also make some introduction on deep learning network, and overview on the commercial institute for inappropriate website classification.

2.1.1 Classification based on URL

Generally, websites can be classified with the contents such as title, description, metatags and link structure. But these factors are unfeasible without visiting, fetching and loading the webpages. Some of papers introduced to classify websites using URL because it is advantageous without processing large contents of webpages.

The uniform resource locator (URL) is a representation used to specify the location of information on the Internet's World Wide Web service. It was originally invented by Tim Berners-Lee to be used as a World Wide Web address. It has now been compiled by the World Wide Web Consortium as the Internet standard RFC1738. Just as there are many ways of accessing a resource, there are also several options for locating it, and the general syntax of URLs simply provides a framework for building new options using protocols other than those already defined in this document. URLs locate resources by providing an abstract identifier for their location. Once the system has located a resource, it may perform a variety of operations on it, which can be abstracted

into the following terms: access, update, replace, and discover properties. In general, only the access method is the only one that needs to be described in any URL scheme.

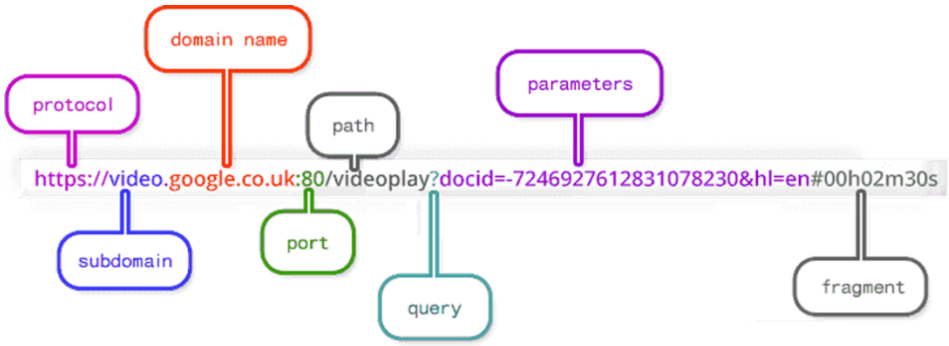


Figure 4. URL Format

Source: [Google]

The URL mainly consists of three parts: the protocol type, the domain name and the path and filename, as shown in **Figure 4**.

Protocol: Specifies the transport protocol used, the following table lists the valid scheme names for the protocol attribute. The most commonly used protocol is HTTP, which is also the most widely used protocol in the WWW. The protocols that can be specified by URL mainly include Http, Https, Ftp, Gopher, Tencent, File, etc.

Domin name: Refers to the Domain Name System (DNS) hostname or IP address of the server hosting the resource. Sometimes, the username and password required to connect to the server can also be included before the hostname (format: username:password).

Port (port number): integer, optional, the default port of the scheme is used when omitted. Various transmission protocols have default port numbers, for example, the default port of http is 80. If omitted when entered, the default port number is used. Sometimes for security or other considerations, the port can be redefined on the server, that is, using a non-standard port number. In this case, the port number cannot be omitted from the URL.

Path (path): a string separated by zero or more "/" symbols, generally used to represent a directory or file address on the host.

Parameters: This is optional for specifying special parameters.

?Query: optional, used to pass parameters to dynamic web pages (such as web pages made using CGI, ISAPI, PHP/JSP/ASP/ASP.NET and other technologies), there can be multiple parameters, use "&" symbol separates the name and value of each parameter with the "=" symbol.

Fragment: Information fragment, string, used to specify the fragment in the network resource. For example, there are multiple noun explanations in a web page, and you can use fragment to directly locate a noun explanation.

A framework provided by Chinnadurai, 2017, stated we can classify the website based on the address bar. The URL's features can be detected and extracted as Table 6 showing.

Table 6. URL Features in Chinnadurai's Model

URL Feature	Detail
Domain of URL	Extract domain name in URL by matching URL phrase.
IP address in URL	Some websites hide their true purpose by using an IP address as a URL.
"@" symbol in URL	When the "@" symbol is used in a URL, the browser ignores anything before the "@" symbol, and the actual address always appears after the "@" symbol.
Length of URL	The length of the URL can also be used as a basis for classification. Generally, longer URLs (such as those longer than 54 characters) have some bad directions.
Depth of URL	By calculating the '/' contained in the URL, the number of sub-pages is obtained, and then the depth of the URL is obtained.
URL redirection	The "//" in the URL means that the page is redirected, and the normal "//" in HTTP and HTTPS should be in the sixth or seventh position.
HTTP/HTTPS	HTTP means Hyper Text Transfer Protocol, information is transmitted in text,but HTTPS is a kind of secure SSL/TLS encrypted transmission protocol. HTTP and HTTPS use totally different connection methods and different ports. HTTP connection is simple and stateless; HTTPS protocol is a network using SSL/TLS+HTTP protocol, which can be used for encrypted transmission and authentication. It is more secure than HTTP protocol.
Shortening URL	Long URLs go through "HTTP redirects" to get a relatively short URL.
Prefix or Suffix '-' in domain	The dash symbol "-" is rarely used in legal URLs.

Source: [Chinnadurai, 2017]

Another framework by Aldwairi & Alsalman, 2012 is also to classify website using URL features. They use lexical features, host-based features and special features.

Table 7. URL Features in Aldwairi & Alsalman's model

URL Feature	Detail
Lexical Feature	Lexical features are properties of the URL itself, excluding the content of the page it points to. URL properties include the length of the top-level domain (TLD), other domains, hostname, URL length, and the number of dots in the URL. Additionally, lexical features include each token in hostname (separated by '.') and tokens in path URLs (separated by '/', '?', '+', '-', '%', '&' , '.', '=' and '_'). These words show some of main content of the website.
Host-based Feature	This kind of information can be achieved on the website WHOIS. Host-based features include IP address, geographic properties, domain name properties, DNS time to live (TTL), DNS A, DNS PTR and DNS MX records. These kinds of feature can help improve the accuracy of detection and classification.
Special Feature	Each of the URL can get JS Enabled/Disabled or HTML Title tag content (<title>, </title>). This kind of information is important for the website topic detection.

Source: [Aldwairi & Alsalman, 2012]

2.1.2 Classification based on text

HTML stands for Hyper Text Markup Language, which is a markup language. It includes a series of labels. Through these tags, the document format on the network can be unified, and the scattered Internet resources can be connected into a logical whole. HTML text is descriptive text composed of HTML commands, which can describe text, graphics, animations, sounds, tables, links, and more. A URL can only point to one HTML page.

In HTML documents, the content of web pages is usually represented by tags, which means that most of the web pages we see are translated from document codes one by one. To get the text information in these web pages, we need to use some libraries to extract useful information.

We can detect the text and classify the website by crawling the HTML file of the website. In this process, we need to clean the tags in HTML, so as to obtain the plain text information of the website.

In Wang & Wang's 2022 paper they introduced a new way to classify text content named Hierarchical Text Classification (HTC in short). HTC categorized text into a set of labels that are organized with structured hierarchy. They use HTC to understand the text content and detect the content labels based on structured hierarchy. Existing methods encode text and label hierarchies

separately, and classify them by mixing their representations, where the hierarchy of all input text remains unchanged. Text encoders can learn to independently generate hierarchical awareness by extracting input text and its positive samples.

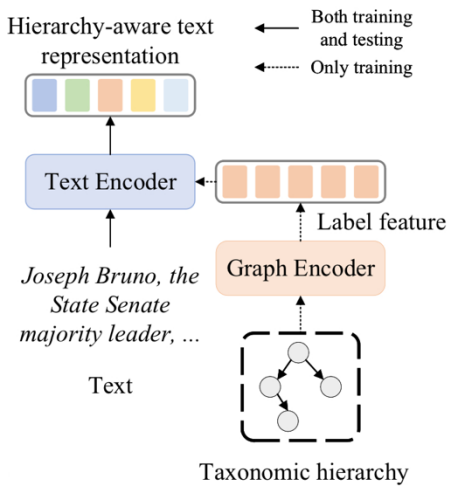


Figure 5. Wang & Wang's HTC structure

Source: [Wang & Wang, 2022]

In the current network environment, there are a lot of content that is not suitable for children, and pornographic content is classified based on age to a large extent. In Glazkova & Egorov's 2020 paper authors introduced an age-based text feature classification model. They obtained a series of characteristics by analyzing the vocabulary and sentence length of people of different ages. They give different weight to age-based vocabulary list, and according to the grade calculated to classify the text is suitable for different ages. As a result, they concluded that their model can effectively screen out words suitable for children of all ages.

2.1.3 Classification based on image

Website classification based on image detection may be the most accurate but also the most time-consuming method. In various literatures, in order to improve the detection efficiency, some use web page screenshots, and some analyze pictures one by one in order to improve the detection accuracy. In this process, we also conducted some research.

In Espinosa-Leal's 2020 paper, they introduce a technique for detecting and classifying bad websites based on ELM (Extreme Learning Machine) website rendering. Extreme Learning Machine (ELM) is a type of machine learning system or method based on Feedforward Neuron Network (FNN), which is suitable for supervised learning and unsupervised learning problems. ELM is regarded as a special type of FNN in the research, or an improvement of FNN and its backpropagation algorithm, which is characterized by the fact that the weights of the hidden layer

nodes are random or artificially given, and do not need to be updated. During the learning process only output weights are calculated. In the websites' screenshots, there are many visual clues, which takes many typical class-related objects.

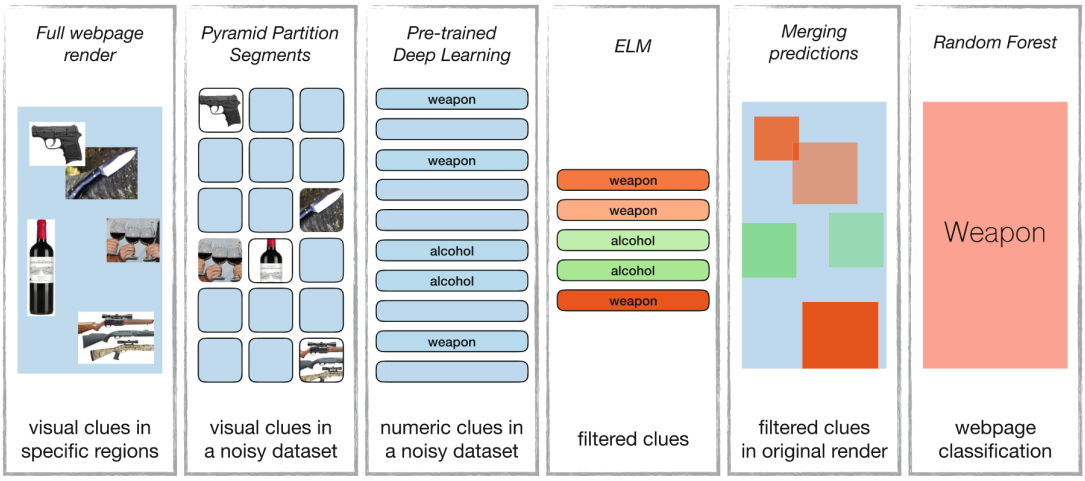


Figure 6. Basic framework of Espinosa-Leal's model

Source: [Espinosa-Leal, 2020]

First intercept a piece of the web page, use the Pyramid Partition Segments (PPS) algorithm to normalize their size. Instead of homogeneously segmented the image, in their approach the windows were built taken as a starting point the top-left corner of the image. They use 2^n pixels (n=5, 6, 7, 8 and 9) of side to square the size during the operation. An ELM model is trained on the characteristic image features from all the samples generated by the Pyramid Partition Segments algorithm.

The goal of ELM is to implicitly learn the background classes of relevant samples and exclude them from the test dataset. Experiments are performed using hyperbolic tangent hidden neurons whose numbers vary from 4 to 32 768. GPU-accelerated toolboxes are used for high-speed calculations. Their algorithm is mainly to detect weapons and other information that is not suitable for children, and it also has certain reference value for the classification of our pornographic websites.

Image features generally refer to regional skin features, as well as baseline features such as color, texture, edge, local shape, and so on. However, due to factors such as face, the accuracy of skin feature recognition is limited. To improve the efficiency of image detection. On the basis of the visual word bag model, Dong and Guo try to merge the text keywords and visual words at the bottom level to expand the vocabulary of the visual word bag. (Dong, Guo & Fu, 2014)

The algorithm is divided into the following parts: Input image files and their web text obtained from the internet, then analyze the image content by detecting the skin area, eliminating small skin area images and extracting local visual features of the image, and then do text analysis by extracting and matching the relevant text with keywords; Finally, features are fused and classified using a SVM classifier to classify the image as adult related or adult independent. The algorithm effectively utilizes the corresponding text information, file names, etc. contained in the web pages, and optimizes the detection results. (See at **Figure 7**)

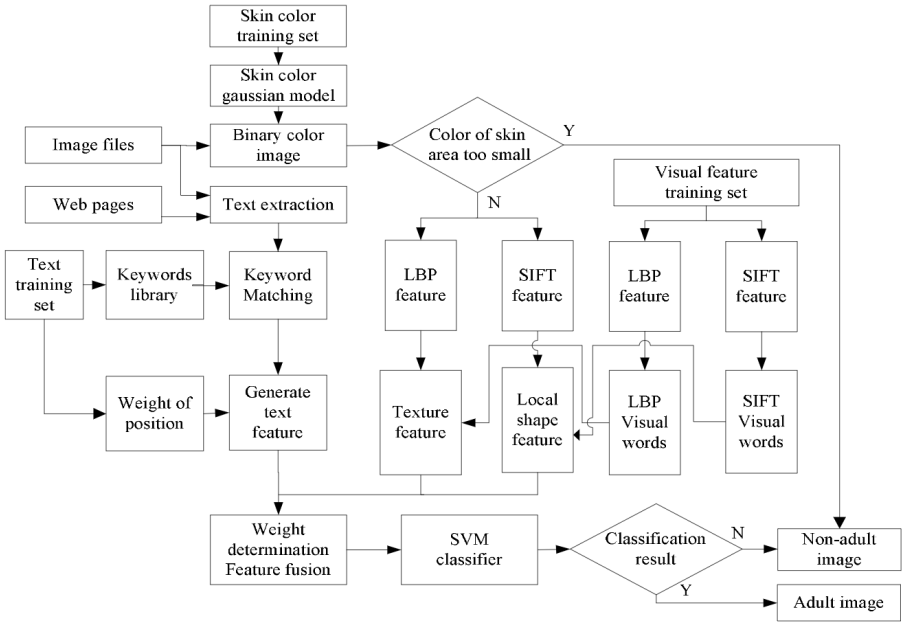


Figure 7. The flow of the adult image detection algorithm

Source: [Dong & Guo, 2014]

However, image detection has high requirements for hardware, and it takes a long time to crawl the pictures in the website. Considering many factors, we finally did not choose the research direction of image detection.

2.1.4 Combination of frameworks

Among the many methods we have mentioned above, we can not only use only one of them, but also combine them to get a model for classifying inappropriate websites.

In the paper by Hammani, 2003, they release kind of framework for classification of website. They combined URL, text and image into one framework like the following structure as **Figure 8**.

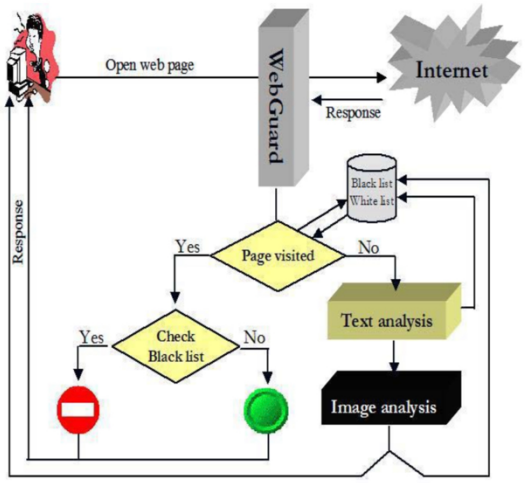


Figure 8. Structure of Hammani's framework

Source: [Hammami, 2003]

They use blacklist to speed up the whole process, and for the strange URL, they use the Text analysis and Image analysis to confirm whether the website content is suitable to be opened, and they set up a Blacklist and Whitelist to store the websites which have been detected. This kind of method use URL Blacklist and Whitelist to filter website fast, and use Text analysis and Image analysis to detect the hole on the Swiss Cheese.

Just URL detection classifying websites in multiple categories also has some models design combined with machine learning and random search, (Shawon & Zuhori, 2018), as shown in the **Table 8**. They use a word-based multi-N-gram model for efficient feature extraction, and provide polynomial distribution for naive Bayesian classifiers under the random search pipeline for hyperparametric optimization, so as to find the best parameters of URL features. Finally, their experimental results are compared with the results of previous research work, and show better results than the existing results. Provided an average F1-score of 87.63%.

Table 8. F1-Score of URL Classification

Category	SVM + all gram [4]	n-gram LM +NB [13]	Multiple n-grams + Random Search +MNB
Adult	87.60%	87.58%	32.03%
Arts	81.90%	82.03%	68.93%
Business	82.90%	82.71%	**86.24%**
Computers	82.50%	82.79%	**95.34%**
Games	86.70%	86.43%	**96.28%**
Health	82.40%	82.49%	**98.37%**
Home	81.00%	81.13%	**95.02%**
Kids	80.00%	81.09%	**81.18%**
News	80.10%	79.01%	**91.26%**
Recreation	79.70%	80.22%	**96.89%**
Reference	84.40%	83.37%	**90.50%**
Science	80.10%	82.52%	**94.83%**
Shopping	83.10%	82.48%	**98.31%**
Society	80.20%	81.66%	**93.12%**
Sports	84.00%	85.30%	**96.21%**
Average/Total	82.44%	82.72%	**87.63%**

Source: [Shawon & Zuhori, 2018]

Through the study of scientific literature, we know that there are mainly URL based, text-based detection and image recognition detection methods in the task of website classification. Combined with our software and hardware strength, we finally chose web site classification based on text detection as our research direction.

2.2 Neural Network Algorithm

In the field of website classification, neural network is also a popular way, and with the development of the times, neural network has also developed many algorithms. Among many algorithms, the following are suitable for the text detection and classification we plan to use.

2.2.1 RNN

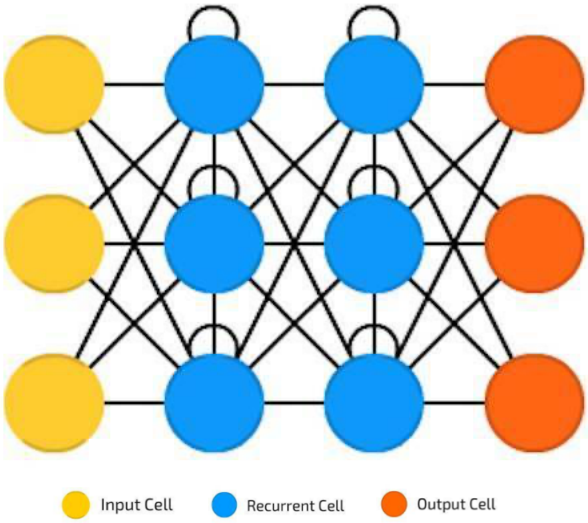


Figure 9. Network structure of RNN

Source: [Author Research]

Recurrent neural networks (RNN) are feedforward neural networks that consider time. They are not stateless, and there is a certain connection between channels through time. Neurons not only receive information from the previous neural network, but also receive information from the previous channel. This means that the order in which you input the neural network and the data used to train the network matters: inputting "milk", "cookies" and inputting "cookies", "milk" will produce different results. The biggest problem with RNNs is vanishing gradients (or exploding gradients), depending on the activation function used. In this case, the information disappears quickly over time, just as information is lost as the depth of the feedforward neural network increases. Intuitively, this isn't a big deal, since they're just weights and not neuron states. But over time, the weights have stored past information. If the weight reaches 0 or 1,000,000, the previous state becomes uninformative.

Convolutional neural networks can be applied to many fields, most forms of data do not have a real timeline (unlike sound, video), but can be represented in sequence form. For a picture or a text string, you can enter one pixel or one character at a time at each time point. Therefore, time-dependent weights can be used to represent information from a second before the sequence,

rather than a few seconds ago. Generally, recurrent neural networks are a good choice for predicting future information or for completing information, such as autocompletion.

2.2.2 LSTM

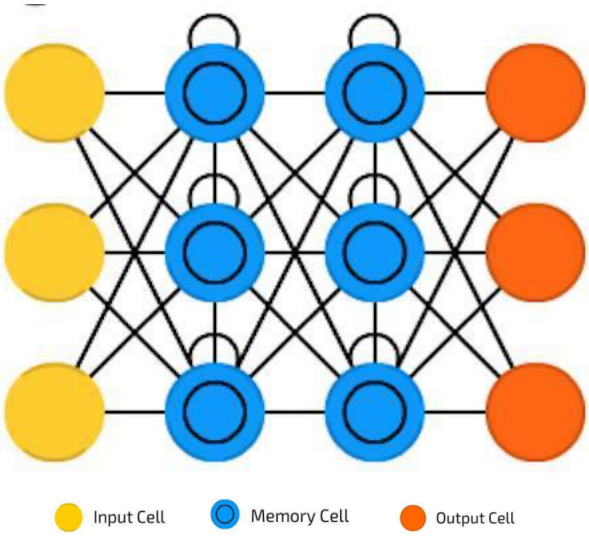


Figure 10. Network structure of LSTM

Source: [Author Research]

Long / short term memory (LSTM) tries to overcome the problem of gradient disappearance or gradient explosion by introducing a gate structure (gate) and a well-defined memory cell (memory cell). Much of this thought was inspired by circuits, not biology. Each neuron has a memory unit and three gate structures: input, output and forget. The function of these gate structures is to protect information by prohibiting or allowing its flow. The input gate structure determines how much information from the previous layer is stored in the current memory cell. The output gate structure does the work of the other end, determining how much information the next layer can learn about this layer. Forgetting the gate structure may seem strange at first, but sometimes forgetting is necessary:

If the web is learning a book and starting a new chapter, it is necessary to forget some of the characters from the previous chapter. Long and short-term memory networks can learn complex sequences, such as writing like Shakespeare, or synthesizing simple music. Notably, each of these gate structures assigns weights to the memory cells in the previous neuron, so generally require more resources to run.

2.2.3 GRU

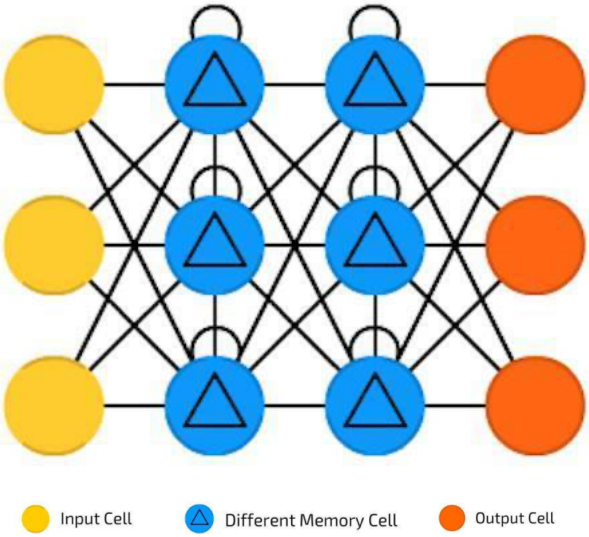


Figure 11. Network structure of GRU

Source: [Author Research]

Gated recurrent units (GRU) are a variant of long and short-term memory networks. The difference is that there is no input gate, output gate, forget gate, it has only one update gate. The update gate determines how much information is retained from the previous state and how much information from the previous layer is retained. This reset gate functions much like the LSTM's forget gate, but its placement is slightly different. It always emits the full state, but has no output gate. For the most part, they function very similarly to LSTMs, with the biggest difference being that GRUs are slightly faster and easier to run (but less expressive). In practice, these tend to cancel each other out, as performance benefits tend to cancel out when you need a larger network for more expressiveness. GRUs may outperform LSTMs without the need for additional expressiveness.

2.3 Commercial software for inappropriate site classification

2.3.1 Commercial institution for website classification—URL-Classification.io

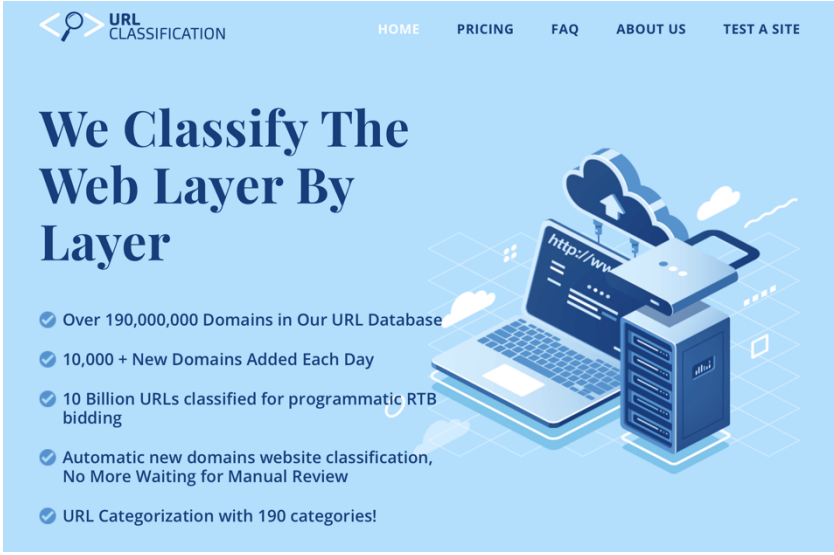


Figure 12. Homepage of URL-Classification.io

Source: [url-classification.io]

URL CLASSIFICATION is a company from Israel. They created a huge URL database including 10 billion URLs classified with 190 categories. The company's database includes 190 000 000 domains with a 10 000+ per day adding speed. They provide URL categorization service for many industries, such as parental control, internet service provider, router application, end point protection and Adtech.

For Adtech industry, which is similar to Quiet Media, URL CLASSIFICATION could achieve the ability to classify individual web pages, which is good for RTB on global sites like a news site, allowing to understand user intent more clearly, and bid accordingly. The accuracy is up to 99.95%, they promise this SLA, over the last four years it's been 100%. And the network is built in such a way that it can compensate for server failure and still provide high-quality service.

The price of this service is a little high, and the charge is constant because the database need to be updated, which is shown as the **Table 9**:

Table 9. Price for service of URL-Classification

Type	Update interval	Total cost	2nd Year updates cost
Active database	One time purchase	20 000 USD	None

Active database	One time purchase+4 updates, once every 3 months	30 000 USD	20 000 USD
Active database	One time purchase+12 updates, once every 1 month	40 000 USD	25 000 USD
Full database	One time purchase	60 000 USD	None
Full database	One time purchase+4 updates, once every 3 months	75 000 USD	30 000 USD
Full database	One time purchase+12 updates, once every 1 month	99 000 USD	40,000 USD

Source: [url-classification.io]

2.3.2 Website Categorization API—WhoisXML API

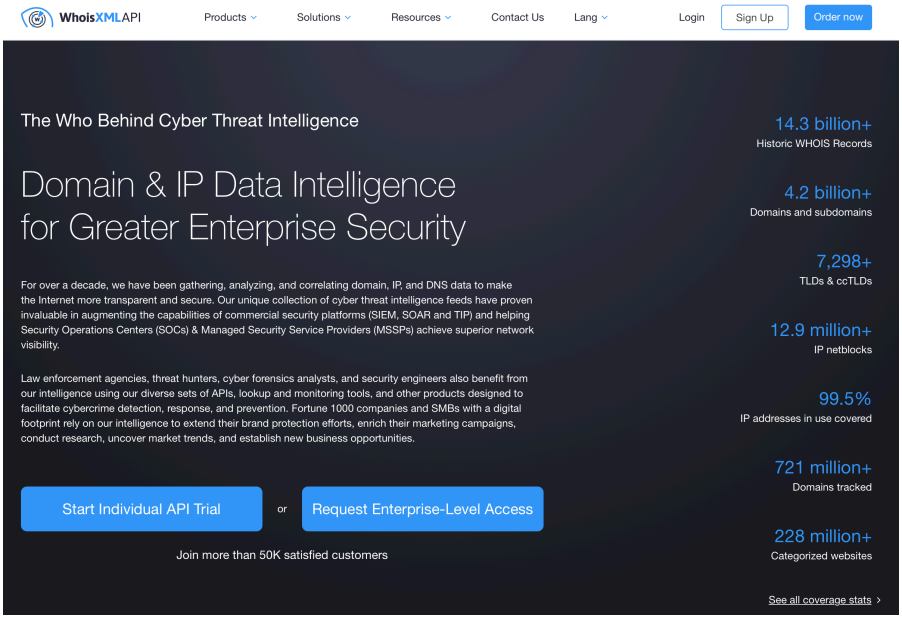


Figure 13. Homepage of WhoisXML API

Source: [https://main.whoisxmlapi.com]

WhoisXML API's website classification tool combines machine learning (ML) and natural language processing (NLP) to scan website text content and MetaTags to classify them. It uses the domain name as input and assigns more than 500 IAB categories and subcategories that are most applicable to each queried site. It also provides confidence score for each category. Basically, the higher the confidence score, the more accurate the classification may be.

By using a combination of advanced ML algorithms and human assistance, their web crawlers parse more than 4 million websites every day, so all data remains fresh and relevant. In

their online website database for commercial use, the domains of websites are divided into 25 categories, which can be purchased according to the needs of the database. For example, if we want to purchase the website category for adult website content, we need to pay $896.08, and then you will get a .csv file and a .json file. As the range of website categories and countries expands, the cost will be higher.

For our whole project, we choose the web site classification based on text detection as the basic research direction, and we will also try to use the neural network algorithm to improve the accuracy of classification. The reason why image detection is not chosen as the research direction is that the training of image detection model takes a long time and the text characteristics in pornographic websites are relatively significant. Existing commercial site categorization platforms are expensive and require frequent access to databases using APIs, so a simple text detection model is sufficient for our lightweight goal of detecting pornographic websites.

Chapter 3. Development of a website classification model for Quiet media

Website classification is an information capture application that offer useful insights which could be basis for many different application domains. Classifying web pages provides an efficient way to Internet using, spam filtering and many other application areas. Search engines like Google and Yandex are making topic-based classification on the website searching in order to get relevant results. Social media platforms like Twitter are using classification model to categorize tweets to recommend then to different types of target customer. In our project, a classification model is needed to automatically extracts information from web pages and categorizes the types of sites to filter adult content sites.

Classification problem is a classic machine learning task which can be solved by supervised learning approaches and unsupervised approaches (Buber & Diri, 2019). Basically, supervised machine learning approaches including two stages: training phase and testing phase. Model should be created to solve the classification problem with number of grouped labeled samples. There are also a lot of experiments on unsupervised approaches, the logic of those approaches is dividing websites into groups according to similarities. Distance calculations are made for determine the similarities of those groups. Hernández et al. (2014) developed an unsupervised web page classification system based on URL.

In our case, we will focus more on supervised machine learning methods because we only have two categories of website need to be classified. Porn sites and sites which are not porn. This chapter will mainly include data collection, preprocessing, model development and evaluation. We will discuss the scale of the primary dataset, types of different data and which kind we are going to use and the reason why we are choosing it, methods which are available for data crawling, the constructure of the data we collected and statistical analysis of them, available methods for transforming collected data. In order to improve the agility and performance of the model, we will also create a corpus to hold adult-related keywords.

As for model building, we will mainly introduce options of algorithms that could be used in the process of building and training the model, metrics of evaluation and model optimization.

The purpose of introducing those methods and procedures is to compare and select the most suitable method for our case in each step, thus finding the optimal model.

3.1 Data collection and preprocessing

Model quality evaluation metrics of input data preparation could be divided into varies steps, including primary dataset preparation, data type selection, data crawling, data cleaning and data preprocessing

3.1.1 Primary dataset introduction

The primary dataset we received is a series of links which was collected by Quiet Media (see **Table 10** blow). This dataset includes the 2 362 541 normal links, 11 982 links in blacklist and 458 022 links in whitelist (the blacklist contain all URLs that have been proven to be pornographic sites, and the whitelist contains URLs that are proven not to be pornographic sites). The dataset to use for training the model includes URL from blacklist and whitelist, because these links are mutually exclusive. It is worth noting that the size of these two lists is not balanced, so this should be considered in future model selection and performance evaluation.

Table 10. Primary dataset

File name	Number of links	Memory usage	Description
all-HTTP-only.txt	2362541	18+MB	A txt file containing a large number of random URLs
blacklist-HTTP-only.txt	11982	93.7+KB	A txt file containing blacklisted porn sites
whitelist-HTTP-only.txt	458022	3.5+MB	A txt file containing not pornographic URLs

Source: [Author Research]

3.1.2 Data type selection

There are various types of information that can be extracted from the URLs, and multiple types of data could be chosen depending on the kinds of classification model:

- **URL features.** The URL features could be extracted and used as keywords to train the classification, especially for domain name.
- **Length of URL.** Calculating the URL's length. Websites will mask the suspicious aspect of URL in address bar by using a long URL. (Chinnadurai,2017)
- **Mega tags.** A well-structured website includes mega tags which contain some information about the website, the most widely used meta tags are keywords, title, description. (Buber & Diri, 2019)

- **Text.** Text was widely used in classification model, mostly because it's the main part of the html including a lot of information.
- **Image link.** Most of the porn sites contain pictures, by analysis the structure of the html, The links of the images can be extracted by means of regular expressions and thus used in image classification models
- **Video link.** Few html of websites contains video links, mostly they just hide the link of the video in different ways, the link of video could be used for doing screenshot and insert into image classification models.

Each type of data has its own usage. In combination with the degree of information contained in the content and the degree of actionability, the text type was mainly selected because of following advantages:

- **High degree of information**: text is the main part of most websites, including website titles, tags, introductions, etc., there is a great value of research and analysis, the use of text type allows the model to better understand the content of the site.
- **Low memory consumption**: Compared to images and videos, text will take up less memory, which is good for later processing of text and the agility of the model. Especially when dealing with large amounts of links.
- **Flexibility**: Text type have low structural impact depending on the site. Many illegal sites or adult sites are not well structured. Therefore, mega tags or the composition of a domain name could have no meaning at all. This may result in inaccurate or less successful collection of information, which is not a problem with the text.

There are also a lot of studies about URL-base websites classification model from scholars (Rathore & Singh, 2019). Additionally, we also extract features base on URL and length of the URL to make some experimental explore.

3.1.3 Data crawling

The web is transforming from document collection to the biggest connected public data space. Large amounts of tools and methods have been developed for collecting data from data-rich websites with efficiency. (Meusel, Mika, & Blanco, 2014)

In our project, data need to be collected from given URLs in order to prepare the training dataset for the model. There are multiple methods could be used in our case. Such as manual collect, website APIs, crawler tools and self-made crawler module in python.

Depending on the type of classification model used, the method and type of information extracted from a web page may also different. (See **Table 11** below)

Table 11. Available methods to collect information from URL

Collect methods	Method Description	Advantage	Disadvantage
Manual collection	Manually enter the website link into browser and copy needed information.	Controllable process, high reliability	Consumption of human and time resources, not suitable for big amount of data set
Website API (Application Programming Interface)	Some companies provide API tools for developers to crawl and analyze data, such as Facebook, Google, etc.	Credible source, easy to operate, high efficiency	Platform limitation, access restrictions
Crawler tools	There are a variety of different well-established crawler software available on the Internet, which can be used to crawl and analyze data in bulk.	Easy to operate, technology maturity, few free for individual usage	Purchase required, function limit, limited content collection.
Self-made crawler module in Python	Make us own crawler module to automatically extract information from websites	Free, customized on demand, high flexibility and feasibility, high efficiency	Requires manual debugging and refinement, highly uncertain.

BeautifulSoup4 Library was used for extract text content from the html, it converts complex HTML documents into a complex tree structure, each node is a Python object, and all objects can be summarized into 4 types:

1. Tag. We can easily get the content of these tags using soup plus tag name, the type of these objects is bs4.element. Tag. Note, however, that it looks for the first matching tag in all content. For Tag, it has two important attributes, name and attrs.

2. Navigable String. Now that we have got the content of the label, to get the text inside the label is also very simple, just use. string.

3. BeautifulSoup. A BeautifulSoup object represents the content of a document. Most of the time, it can be regarded as a Tag object, which is a special Tag, and we can get its type, name, and properties separately.

4. Comment. The Comment object is a special type of Navigable String object whose output does not include comment symbols.

After considering flexibility and feasibility, we opted for self-made crawler module combined with beautiful soup library. The reason are as follows:

1) Avoiding copyright issues: most of the current tool and software are only free for individual usage, considering the project is business related, a self-made crawler module could avoid the copyright problem and potentially save the budget.

2) Efficiency: self-made crawler could be used without limitation of the other software environment including usage time limit, traffic limit, memory limit. At the same time, self-made crawler module can save a lot of time compared to manually collecting.

3) Flexibility: Due to the complexity of given websites and the different types of data required, self-made crawler could provide customize features to suit needs of project. For instance, most of the porn websites are not registered through official channels, and the structure of the html (HyperText, Markup Language) also different.

The working process of the crawler are as follows (See Figure 11 below): First we make request to each URL, then used try and except blocks from python to get all the URLs that can be accessed and get the html from them, then we save each html as a html file and name it by URL. For URL which are not accessible, just ignore it and save the name of URL as well. The next step is to get content which we need from URLs and files. By analysis the URL list, features and length can be collected. Then extract text by using the module beautiful soup, the text data and language type of the website will also be collected in the DataFrame. In the end, save the DataFrame into a csv file. (See **Figure 14** below)

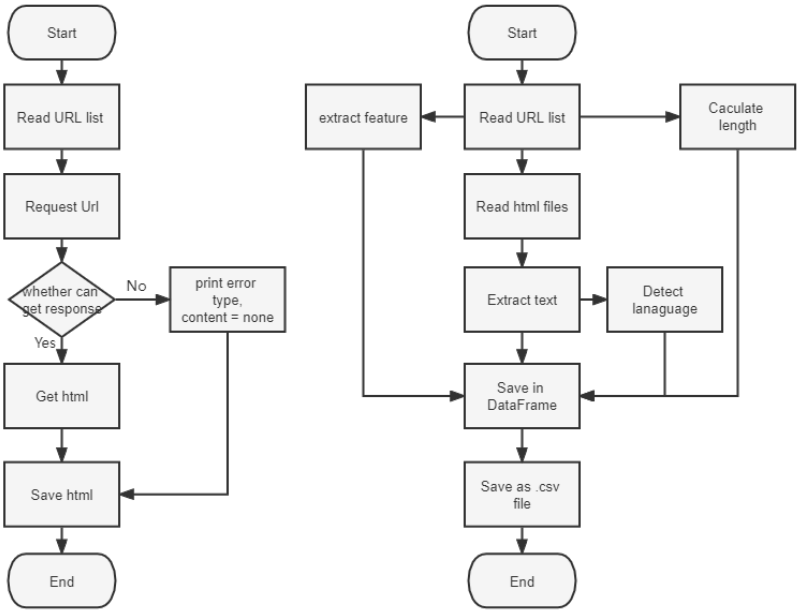


Figure 14. Processes of data crawling

Source: [Author Research]

3.1.4 Data cleaning and integration

In order to get the data which could be used for model training, the process of data cleaning and preprocessing are necessary. Data cleaning is commonly used for fixing or removing incorrect, corrupted, wrong formatted, duplicate, or incomplete data in a dataset. There are different solutions for different problems. For incomplete data, available attempt to complete the data, either by re-collecting it, completing it by context, or ignoring it. This helps to ensure the integrity of the data. For duplicate data, the most frequently used approach is to delete duplicate records and keep only one. This helps to reduce the complexity of the data and avoid redundant calculations. For incorrect data, most of them are data that do not conform to common sense, for instance, a negative age, a birth date greater than the date of the day, or an outlier. The most frequently used method can be to set specific rules to remove incorrect data, ignore or reject outliers according to their reasonableness. In our cases, we deleted websites without contents or in inaccessible pages, mark up pages in different languages according to the language type of the text in the page and remove pages whose content cannot be detected by language type. After these operations, which facilitate the reduction of data complexity, a clean dataset can be obtained and used for subsequent model training.

After data cleaning, all the text should be preprocessed. The purpose of this step is to reduce the complexity and repetitiveness of the text, so that it is better recognized, and to remove special symbols and line breaks. We used the method replace, to find out all the line breaks and special symbols like '\n','\t', 'r' and replace them into blank space. It`s also available to use regular regression to select and replace some meaningless words by rules, like by setting the rule:' [^a-яA-Яa-zA-Z] +', words only containing English and Russian letters will be selected and all uppercase letters will become lowercase, other special letters or numbers would be dropped automatically. (See **Figure 15** below)

text	language	text_processed
1000 Porn Home Popular Latest Longest Englishx...	en	porn home popular latest longest englishx engl...
1001 XXX Home Popular Latest Longest Englishx ...	en	xxx home popular latest longest englishx engli...
101 Erotic Girls | beautiful girls, barely leg...	en	erotic girls beautiful girls barely legal mode...
123 Porn Tube Home Popular Latest Longest Engl...	en	porn tube home popular latest longest englishx...
Teen Porn and Erotic Pics, Girls Ass Fuck @ 18...	en	teen porn and erotic pics girls ass fuck porn ...
...	...	...
XNXX, ZOO Porn, XXX Tube, Free Bistiality Vide...	en	xnxx zoo porn xxx tube free bistiality videos ...
XVIDEOS, ZOO Tube ,Free ZOO PORN, XNXX ZOO Vid...	en	xvideos zoo tube free zoo porn xnxx zoo videos...
Free porn @ Zoya Porn ZoyaPorn English Ελληνικ...	en	free porn zoya porn zoyaporn english galego vi...
Zumba XXX ZumbaXXX English English Čeština Dan...	en	zumba xxx zumbaxxx english english e tina dans...
Free Porn Pics and Best Sex Photos milfpics.mo...	en	free porn pics and best sex photos milfpics mo...

Figure 15. Text before and after processed

Source: [Author Research]

Since our dataset contains both English and Russian languages and the two languages have different grammars, there is a module called ‘pymorphy2’ that can be applied. Pymorphy2 is a morphological analyzer which could be used for Russian languages. By using morphological analysis, we can analysis the internal structure of words. Especially for languages with rich morphology, for instance, Russian language (Korobov,2015). By using pymorphy2 module, it’s possible to figure out the lexical nature of the word, whether it is a verb or a noun or other natures. By analyzing the lexical properties of each word in the text, we are able to extract and remove words that have no real meaning, such as prepositions like ‘в’, conjunctions like ‘и’, and exclamations like ‘ой’ or other particles like ‘бы, же, лишь’. Another function of this module is to revert words that have been conjugated to their original form. For example, the words like ‘куплю’ or ‘купил’ verb could be changed to their original format which is ‘купить’. (See **Table 12** below) In this case, The repetition of sentence meanings in the text in the dataset can be reduced, the size of the text can be reduced, and the complexity is reduced, which is beneficial for training the model.

Table 12. Text processing by using 'Pymorphy2'

производство строительных лесов продажа рамных...	производство строительный лес продажа рамный л...
анекдоты на сайте a ru все темы новые анекдоты...	анекдот сайт a ru тема новый анекдот сайт a ru...
русскоязычный центр интернет форма регистрации...	русскоязычный центр интернет форма регистрация...
сайт для корпоративных клиентов сервис телефон...	сайт корпоративный клиент сервис телефон главн...
территория права перейти к содержанию территор...	территория право перейти содержание территория...
...	...
оф сайт diy s and advice обзоры плейлисты лайф...	оф сайт diy s and advice обзор плейлист лайфха...
продажа столов стульев барных стульев челябинс...	продажа стол стул барный стул челябинск вход р...
томатов большая коллекция лучших редких сортов...	томат больший коллекция хороший редкий сорт то...
традиций со всего мира traditions com главное ...	традиция весь мир traditions com главное меню ...
электроинструмент и строительное оборудование ...	электроинструмент строительный оборудование ро...

Source: [Author Research]

3.1.5 Data construction

By coding and improving the crawler module, we crawled the textual content of the website for the given dataset and saved them into csv files, including 7 878 data raw from blacklist, and 246 965 data raw from whitelist. After that according to the language type of the text, the following files was saved. (See **Table 13** below)

Table 13. Information of Data files

File name	length	Description	Memory usage
dfwhite_en.csv	9 719	The file contains URLs of the normal English web pages, domain name, length of the URL, the original text and the pre-processed text	391MB
dfwhite_ru.csv	230 488	The file contains URLs of the normal Russian web pages, domain name, length of the URL, the original text and the pre-processed text	2.19GB

dfblack_en.csv	4 406	The file contains URLs of the porn English web pages, domain name, length of the URL, the original text and the pre-processed text	62.2MB
dfblack_ru.csv	2 832	The file contains URLs of the porn Russian web pages, domain name, length of the URL, the original text and the pre-processed text	49.8MB
dftrain_en.csv	14 125	The file contains information about all the English websites that will be used as input model data (See Table 5 below)	142MB
dftrain_ru.csv	233 320	The file contains information about all the Russian websites that will be used as input model data	2.38GB

Source: [Author Research]

Columns of each files include the primary URL, extended URL, text which collected by crawler module, language of the text, preprocessed text and marker. (See **Table 14** below)

Table 14. Column names and description

Column names	Description
Raw	Primary links
URL	'Http://'+ primary links
URL_dmoain	Domain name of each website
URL_length	The length of each URLs
Text	Texts collected from websites by crawler module
Proc	Preprocessed clean text
Language	Language types of the text
Marker(Y)	Y=1 the website is a porn site, Y=0 the website is not a porn site.

Source: [Author Research]

In order to have better understanding with data, some statistical analysis is needed.

(1) URL: Due to the instability of the site, many of the original links are inaccessible, or the web server refuses access. Therefore, only about 65% of the blacklisted sites and 53% of the whitelisted sites are collected. (See **Figure 16** below) According to **Figure 16**, two feature of the dataset shows up. The first one is that two list are not balanced, another is the ratio of readable website in each list is not high, which should be notice further.

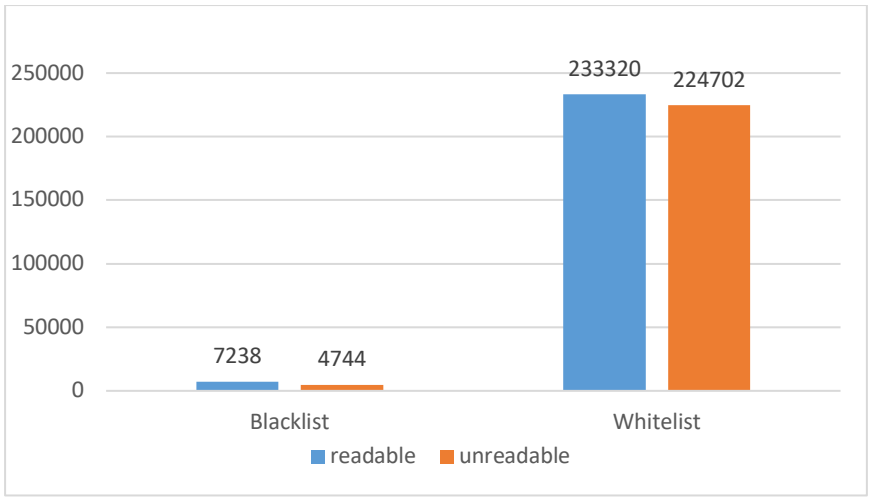


Figure 16. Number of readable and unreadable links in raw dataset

Source: [Author Research]

(2) Language type: As for language type of texts, in all 7 878 readable links from blacklist, 4 406 websites are in by English and 2 832 of them are in Russian. For whitelist, 230 488 are in Russian and 9 719 of them are in English. (See **Figure 17** below) As for whitelist, 230 488 of websites from it are in Russian and 9 719 of them are in English. (See **Figure 18** below)

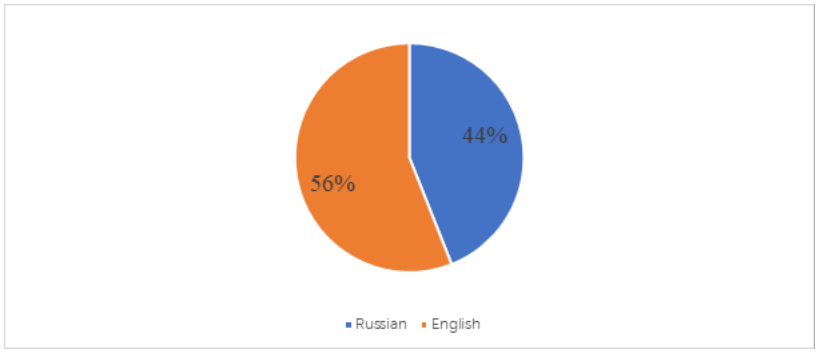


Figure 17. Type of website text language in blacklist

Source: [Author Research]

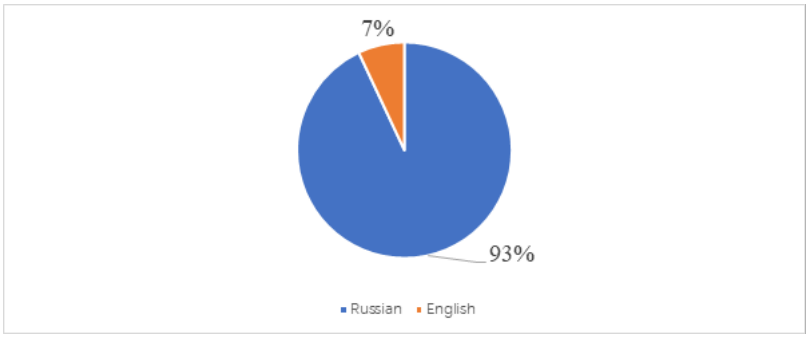


Figure 18. Type of website text language from whitelist

Source: [Author Research]

According to the distribution of languages, few things should be noticed:

1. Language distribution in whitelist is also not balanced.
2. Two models need to be trained in order to classify different language types of websites.

(3) URL length: The length of URLs is one of the features we extracted from the dataset, to have better understanding on that, we divided it into 4 parts: 'URL length in English whitelist', 'URL length in English blacklist', 'URL length in Russian whitelist', 'URL length in Russian blacklist'. Different distribution appeared in different part, here are the brief information about the length in each part. (See **Table 15** below)

Table 15. URL length in different datasets

Feature name	Counts	Mean	Std	Min	Median	Max
URL length in English whitelist	9719	21.16	4.53	12	21	60
URL length in English blacklist	4406	23.97	3.99	13	24	55
URL length in Russian whitelist	230488	20.80	5.54	11	20	78
URL length in Russian blacklist	2832	28.04	8.26	13	26	59

Source: [Author Research]

When we talk about the distribution of lengths, placing the data in a chart could shows the difference between the various data lists. (See **Figure 19** below)

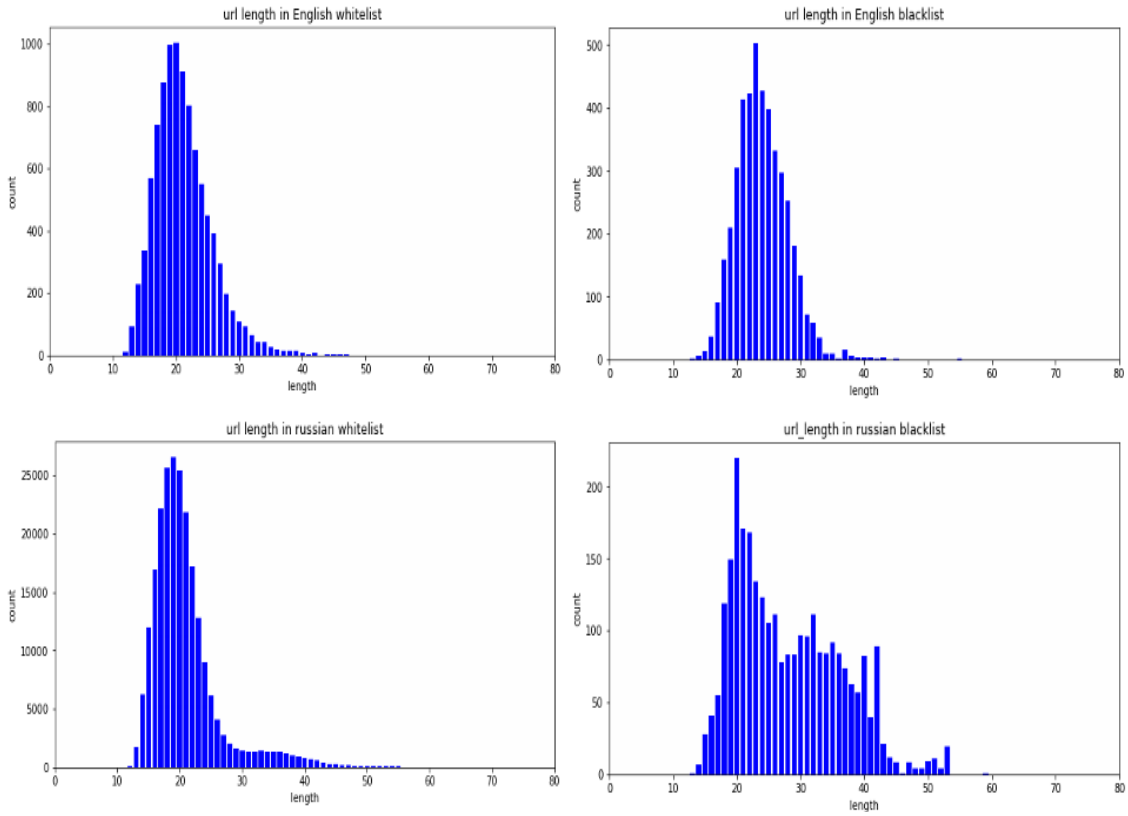


Figure 19. Distribution of varies data lists

Source: [Author Research]

According to **Figure 19** we can see that most of the sites show a normal distribution in length, but the degree of blacklisting of sites in Russian is relatively unevenly distributed. To go further, we fitted the data and plotted the fitted curves. (See **Figure 20** below) Through the chart we can clearly see that the distribution of URL length of ordinary websites show generally normal distribution, while the distribution of URLs in the blacklist has a slight incomprehension and is not uniformly distributed, indicating that the length may be used as a useful feature to distinguish between different types of URLs in the classification.

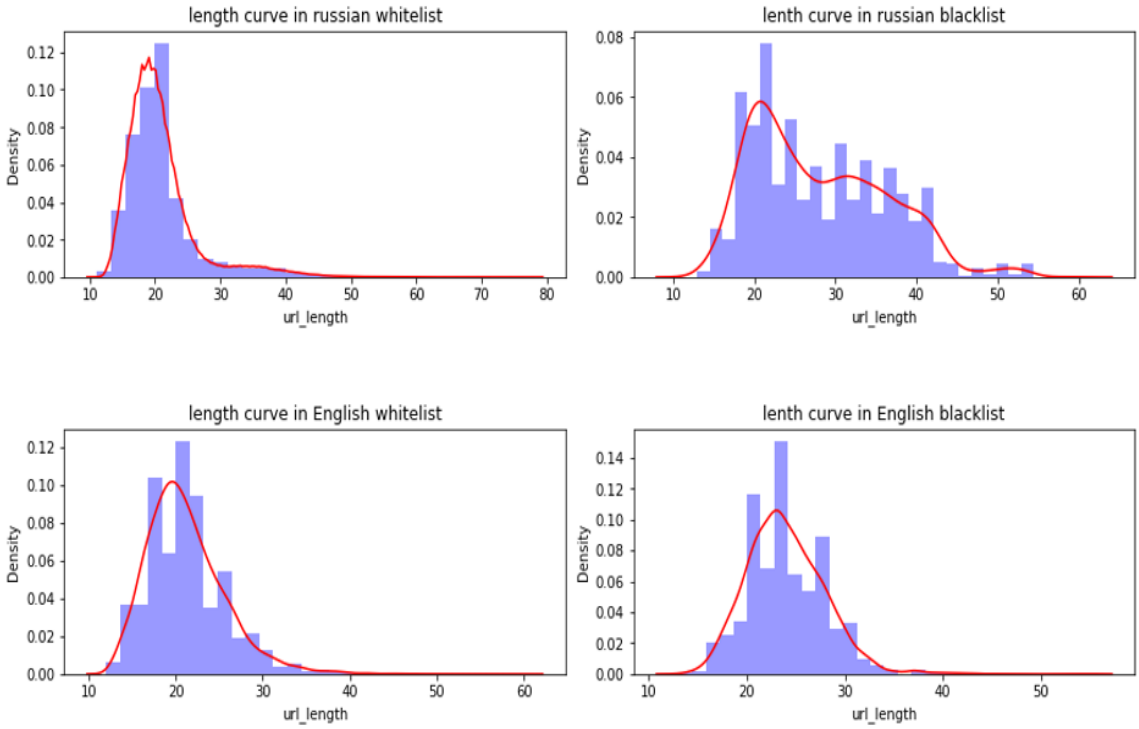


Figure 20. Fitted curve in different data lists

Source: [Author Research]

3.1.6 Feature engineering and data transformation

After preparing the dataset, the next step would be select features as the input of the model. The main idea of this steps is to extract suitable feature from information we collected from links.

Various feature selection methods could be used to drop irrelevant features and group the reduced features as a new subset (Shabuddin et al., 2020). Hybrid approach is the combineation of filter approach and wrapper approach

According to the structure of the current dataset, wrapper approach would be suitable for feature selection and can decrease the complicity and increase flexibility. Wrapper approach choose few subsets of features and then evaluated them by using classifiers. (lqbal et al., 2020).

Second, convert text features into numeric vectors and use them as input to the model.

Next step is to transform the text into a form which could be recognized by computer, text document can be represented either in the form of binary data, when we use the presence or absence of a word in the document in order to create a binary vector. In this situation, it is possible to directly use a series of categorical data clustering algorithms on the binary representation. A more enhanced representation would include refined weighting methods based on the frequencies

of the individual words in the document as well as frequencies of words in an entire collection, for instance, frequency/inverse document frequency (TF-IDF). Quantitative data clustering algorithms can be used in conjunction with these frequencies (Charu & Zhai, 2015).

In text classification, the document may partially match many categories. We need to find the optimize matching category for the document. TF-IDF approach is commonly used to weigh each word in the text document according to how unique it is. In other words, the TF-IDF approach captures the relevancy among words, text documents and particular categories (Zhang et al., 2005).

There are some limitations of TF-IDF algorithm that needs to be addressed. The major constraint of TF-IDF is, the algorithm cannot identify the words even with a slight change in its tense, for example, the algorithm will treat “go” and “goes” as two different independent words, Due to this limitation, when TFIDF algorithm is applied, sometimes it gives some unexpected results. Another limitation of TF-IDF is, it cannot check the semantic of the text in documents and due to this fact, it is only useful until lexical level (Qaiser, 2018).

There are two parameters which should be set before applying the method in python, which are ‘max_df’ and ‘min_df’. These two parameters are used to downsize the text to avoid simple meaningless words like prepositions and conjunctions.

TF-IDF can help in the field of decreasing the complexity of the model and transform words into number which could be used to train the model. (See **Figure 21** below)

	url_domain	url_length	all	copyright	policy	privacy	reserved	rights	ru
1500	bejunmehta	21	0.082571	0.044722	0.000000	0.000000	0.040097	0.039886	0.00000
1427	batteryshop	22	0.183850	0.298732	0.266903	0.275336	0.267837	0.266425	0.00000
3485	eastindiashipping	27	0.150021	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000
921	arsenal	19	0.070509	0.000000	0.102360	0.105594	0.102719	0.102177	0.00000
2850	dailycodebuffer	26	0.028416	0.092346	0.082507	0.085113	0.000000	0.000000	0.00000
2951	deborahspiro	23	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000
3701	elpackaging	22	0.060017	0.019504	0.034852	0.035953	0.017487	0.017395	0.00000
12234	teenwetporn	22	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000
5095	happyclub	19	0.000000	0.000000	0.102545	0.105785	0.000000	0.000000	0.27689
538	amatures	19	0.000000	0.000000	0.102118	0.105345	0.000000	0.000000	0.00000
9295	windowsbulletin	26	0.026310	0.000000	0.012732	0.013134	0.000000	0.000000	0.00000
4273	floralinq	20	0.046592	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000
7349	mauropoluzzi	22	0.474154	0.000000	0.000000	0.101442	0.098680	0.098159	0.00000

Figure 21. Sample result of TF-IDF

Source: [Author Research]

3.1.7 Blacklist of words

In order to get better understanding on the text we collected and preprocessed. General insight should be given from analyzing. The basic idea of this part is to find out the most common words which could be potentially influence the type of the website, which is the predicted result from the model. The methods which could be used are word cloud and building blacklist of word.

1) Word cloud

Word cloud is a visualization description of the keywords, it used to generalize the contents from web or text. It used as a means to provide overview by downsizing content to words which shows up frequently in the text. (Heimerl et al., 2014)

In our case, word cloud was used to get insight from the whole text. For instance, the most frequently occurring words in all pages. The way we apply is to place all the text on the site in a string and use word cloud to display the most frequently occurring words.

2) Sensitive words selection

Although using word cloud can bring us insight but there are still meaningless or irrelevant words shown from the result. For instance, the meaning of the words like ‘best’, ’model’, ’added’ which are shown in the result, are irrelevant with porn websites. So, we need to find out a way to narrow the selection, remove prepositions and conjunctions, then manually select the words that

best match the results and create a blacklist based on those words. TF-IDF could be an efficient way to solve the problem. The Implementation method is to first use TF-IDF on the text we collected in python. Get the name of the columns and manually select the words. (See **Table 16** below)

Table 16. Blacklist of sensitive words

Blacklist of sensitive words				
A*al	Bd*m	Blo*job	Bo*bs	Bu*ty
Co*k	Crea*ie	C*m	Dee*throat	Di*k
Di*do	Do**ystyle	Er*tic	Fing*ring	F*ck
Fuc*ed	Han*job	Ho*ny	Mast*rbation	Na*ed
Nip*les	Pi*sing	Pant*es	Po*n	Po*no
Po*nstar	Pu*sy	S*x	Sl*t	Su*king
Thre*some	T*ts	Vi*gin	xxx	Sto*kings

Source: [Author Research]

3.1.8 Data problem

Because of the amount and scale of the links, also most of the websites are not well-structured, problems could be happened in every step during data preparation, those problems may cause the limitation of method and model usage, increasing the time consumption and resources consumption. The problems during those procedures and solutions are presented by the following table. (See **Table 17** below)

Table 17. Data problems and solutions

Stage	Problem	Problem description	Solution
Website crawling	Website cannot be accessed	Some websites can`t be accessed due to varies reason, for instance the websites already closed, server refuse to access. unavailable for legal reason, those problems may cause the coding stop running, report issues and influence the progress especially in the middle of the large list.	1.Add header 'User agent' could mask the module as a browser accessing the website, this can solve the refuse accessing issue, especially the site is pre-set with an anti-crawler mechanism. 2.Using 'Try' and 'Except' during the process of collecting the data could avoid most of the issue

			reporting. It not automatically identifies the type of the error and the skip it.
	Timeout	Clawer may stick in some websites during data collection, especially when the site is not well-structured and not officially registered, low performance website server also may cause this problem	Set parameter' Time out' could limit the time which the module will wait. It can avoid the time out problem and skip the website which are not easily accessed. It also reduces the time consumption of module operation
	Empty websites	Websites could be empty for reasons such as being out of operation or being shut down, which cause the collected data to be empty.	Drop Null values, this can save the memory and reduce time wastage
Language detection	Meaningless Text	After collected the text, some issues may be caused like the website itself only contain special characters such as a series of question mark. This may cause the code could not detection the language type of the website and stop running because only str are needed.	Using 'try' and 'except' could perfectly solve this problem, it`s also can mark the unreadable text in order to drop them in next step. It helps to keep the data cleaner and simpler.
	Unreadable text	Some texts are unreadable or empty	
Data preprocessing	Different language type	During data preprocessing, different language types show up which means different process are needed also in will influence the feature extraction and model training in the following step	Using different vocabulary and stop words to process different language type of texts. Because most of the text are written in Russian and English, two different process method and classification model are needed
	Unbalanced dataset	According to the statistical analysis of the input data, unbalanced distribution of website types and language types of issues was shown. Unbalanced dataset may influence the performance of the model and evaluation.	Dataset extension, collect data from other resources may reducing imbalances in the data set. Be careful in each step, try different method to avoid problem for instance, by using cross-validation or stratified k-fold.

Source: [Author Research]

3.2 Development of a model for website classification

After the preparation of the training dataset, the following step should be modeling building, model training and model evaluation. Basic idea of this step is to first split the dataset as training dataset and testing dataset. First, we use algorithms to classify the training dataset and train the model. After that, multiple metrics would be used to evaluate the performance of the model and based on the result, we can choose the optimize method.

In this stage, cross validation would be used for dataset splitting. The algorithms like logistic regression, random forest and neural network algorithms would be used as classifier to help us building the model and evaluate them.

3.2.1 Data split

It is a methodological mistake that learning the parameter of the predict function and testing them on the same data. Problems like overfitting would show up. In order to avoid it, the step which data split would be necessary. There are several methods for split the data, for instance, randomly split the dataset into two parts by a given ratio, cross-validation etc.

Compare with random split, CV (Cross validation) could avoid the risk of overfitting and improve the universality of the model which means to make sure the model can perform well even in complex application situations.

In our case, different datasets are not balanced, so the method which called stratified k folder cross validation could be used to split the dataset, into training dataset and test dataset.

This cross-validation object is an optimized version of K-Fold that returns stratified folds. The folds are created by saving the percentage of samples for each class.

3.2.2 Algorithms for web site classification

Next would be the classifier selection. Multiple classifiers could be used in this step. In order to find the optimal solution, we need to put the same training dataset into different classifiers. By comparing the score and consumption of time, choose the optimal classifier which is suitable for the project. The classifier that will be used for model training are logistic regression, random forest, recurrent neural network (RNN), long short-term memory (LSTM).

Logistic regression is one of the commonly used algorithms for classification model, this algorithm does not require over-set parameters and is relatively simple, which reduces the complexity of the model while improving the response time. But the disadvantage is that it does not perform well for some complex cases

Random forest is relatively complex compared to logistic, it can adapt to responsible cases and will not have overfitting problems, the parameters to be set is mainly the number of estimators.

RNN and LSTM are two Neural Network Solutions for text classification. LSTM solves the problem of gradient dispersion compared with RNN, and is widely used in language recognition, text recognition, etc.

3.2.3 Model quality evaluation metrics

In Machine Learning, performance measurement is an essential task. Evaluating model provide the insight of the performance of the model. By calculating the score of the model when using the test dataset, the quality of the model can be estimated. Various scoring methods can be used at this stage, such as AUC-ROC curve, confusion matrix, precision, recall and F measure.

Confusion matrix is a classic method which could be used to present the summary of prediction results on classification problems. Precision and recall are two important metrics which could be calculate from the results of confusion matrix. Precision is the fraction of instances among the retrieved instances, recall is the fraction of relevant instances that were returned F measure is the harmonic mean of precision and recall. F measure is calculated as: F-Measure = 2*(Precision * Recall) / (Precision + Recall). F-Measure create the solution to combine both precision and recall into a single measure that catch both properties.

When we need to check or visualize the performance of the multi-class classification problem, we use the AUC (Area Under the Curve) ROC (Receiver Operating Characteristics) curve. It is one of the most important evaluation metrics for checking performance of any classification model.

AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model can distinguishing between classes. Higher the AUC, the better performance the model is at predicting 0 classes as 0 and 1 classes as 1.(Narkhede, 2018)

Advantage of AUC-ROC are as follows:

1) Insensitive to category imbalance

2) Applicable for evaluating classifier overall performance.

3) Does not change significantly with the change in the proportion of positive and negative samples in the sample

After analyzed the distribution of the dataset, we learned that the dataset was not balanced (see 3.1.4). AUC-ROC can avoid the problem in order to have better measurement of model performance which could be the main metric for evaluating our model.

In the process of evaluation, datasets would be splitted into multiple folds by using stratified K-folder method. The score will be calculated by averaging the scores of multiple folds

3.2.4 Text based website classification models evaluation

Evaluation is one of the most important processes when building the model, not only it can provide insight of the performance of the model, but also can help to select the optimize feature, we used control variates methods to check the performance of the model in different combination of the features and the same classifier by using different metrics. After finding out the best combination of the feature, the result could be used in the step of algorithm selection.

In the process of evaluation, wrapper approach is used to choose various subset of features are first identified then evaluated using the same classifiers. By combining different features and using logistic regression for training and testing, we record the scores of AUC-ROC, recall, precision, F1-measure, and the time used for training and testing the model for different combinations. (See **Table 18** below) According to the result, we chose to use preprocessed text data, which is 'proc', as the main feature for subsequent model design and training, based on a combination of measurement time and model complexity and AUC-ROC score.

Table 18. Feature evaluation using wrapper approach

Feature	Classifier	AUC-ROC	Recall	Precision	F1 measure	Time, sec
url_domain	Logistic regression	0.64	0.31	0.87	0.46	2
url_length		0.53	0.49	0.13	0.2	4
proc		0.96	0.93	0.99	0.96	13.7
Proc, url_length		0.96	0.93	0.94	0.93	13.9

Source: [Author Research]

The next step is to select the optimal TF-IDF parameters and the optimal model. The experimental approach is to design different combinations of classification algorithms and TF-IDF parameters, bring them into the model for training and testing, and derive the optimal combination by using multiple metrics. (See **Table 19** below) based on the recorded result of the experiment, LSTM model is the optimal model in terms of classification performance.

Table 19. Part of model evaluation base on test datasets

Algorithm	TF-IDF values		AUC-ROC	Recall	Precision	F1 measure	Time, sec
	Max_df	Min_df					
Logistic regression	.8	.1	0.964	0.932	0.994	0.962	27.6
	.7	.05	0.962	0.925	0.992	0.958	34.5
	.6	.01	0.963	0.925	0.994	0.958	48.1
Random forest	.8	.1	0.975	0.957	0.979	0.968	38.4
	.7	.05	0.975	0.957	0.981	0.969	47.6
	.6	.01	0.976	0.957	0.982	0.977	340
LSTM	8	1	0.989	0.971	0.971	0.971	225
	7	05	0.986	0.974	0.982	0.978	238
	6	.01	0.979	0.969	0.972	0.971	280

Source: [Author Research]

To verify the validity of the model, we applied the model to the same sample containing 200 URLs, while adding manual classification for comparison. The accuracy and the time spent on classification were used as the evaluation criteria. (See **Table 20** below) Due to the complexity of the network situation, many pornographic sites are illegally set up and changed more frequently, and the situation of different URLs is always updated. In the process of manual classification, problem was found like, some URLs are inaccessible, or the domain name is being sold, or there is no content, and all these problems can potentially affect the classification results.

Table 20. Comparison between manual and ML methods

Method	Test sample	Correct prediction	Time, sec
Manual	200	196	1506
Logistic regression		194	750
Random forest		197	736
LSTM		198	738

Source: [Author Research]

In addition, we analyzed the results and process of the content and found that there are many factors that affect the performance of the model, especially since some websites have relatively little text content, so there is only a slight effect on the results after using TF-IDF, and

there are many unavoidable factors including inaccessibility of the website itself, sale of the website, change of the website domain name, data not correctly labeled, etc. Models can be optimized in many places, and there are many options. In future applications, we can choose to train new models with monthly updates and use transfer learning and data augmentation to enhance the flexibility and generalizability of the models.

Based on the experimental and evaluation results, the optimize combination of the features would be the preprocessed texts, the change of parameters of TF-IDF only showed slightly change on the evaluation results. We decided to use the LSTM method as the optimize solution to the project. In the future application and improvement process of the model, those are several methods are appliable. For instance, transfer learning and data augmentation could be used to improve the flexibility of the model. Also, monthly update would be necessary for the company to keep the quality of the predict results.

Chapter 4. Business analysis of the developed model deployment

As a lightweight pornographic web page classification tool, we should evaluate the whole product project from business perspective. In terms of market demand, we need to do a market analysis. We need to find the positioning of products in the known market, so as to formulate targeted promotion strategies. In terms of cost structure, we need to build a cost model to deduce the cost of project operation. In terms of business model, we need to analyze the financial model and draw a specific feasible scheme to confirm whether our products have a profitable market or meet the market demand.

4.1 Market analysis

4.1.1 P/MF model

In 2007, VC guru Mark Anderson coined the concept of P/MF (Product / Market Fit) and defined it as "the ability to meet a market with a product in a good market" in pursuit of matching the product to the market. (See **Figure 22** below)

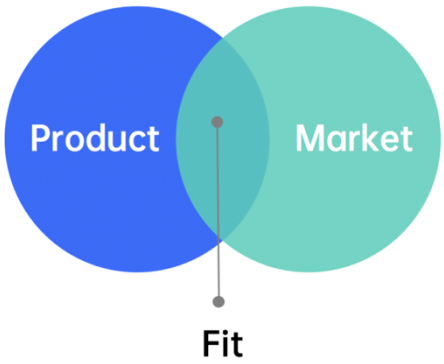


Figure 22. P/MF model concept

Source: [Author Research]

There are 3 types in the P/MF model.

- The first type of P/MF: Satisfy an existing market with a better product experience. The demand is already there but need to experience a better product. The focus is on a very good user experience and a lot of marketing and promotional investment.
- The second type of P/MF: Provide a product to satisfy an existing but partially unsatisfied market. Some of the user's needs are not being met and the segmentation

of the user's needs are being met. The focus is on attracting segmented target customers with a more refined marketing and promotion strategy.

- The third type of P/MF: Provide a product to create a new market. Before the product is created, users do not know they need the product, so the demand does not exist, and the market does not exist. At this point, the product is used to create a new market.

4.1.2 Product review

Our model is targeted as a lightweight tool for classifying pornographic web pages. Our market segment is very clear, targeting clients in the Internet advertising industry who need to avoid pornographic content. For example, if a food brand is associated with a pornographic website when advertising, it will cause a certain degree of consumer churn. On the other hand, we can also provide targeted website advertising to customers whose products are positioned as adult products. For example, erotic product companies are more likely to advertise on such websites because the usage scenarios and target groups are similar. The price of our products is certainly far lower than that of similar products in the market (see at **Table 9**), which can also enable our customers to meet their segmentation needs with less money.

It is also a big task to promote and market our lightweight tools. We mainly divided into several paths. First, we should do a good reputation in customers and conduct in-depth research on the needs of a customer, so as to harvest a stable source of customers in the Internet advertising industry. The second is the customer rebate. Those who use the service for two consecutive years can be given a certain discount price, so as to stabilize the long-term customer source. Last but not least, we should continue to optimize our products, constantly make progress in a better direction, with higher accuracy and a wider range of applications (such as multilingual applications).

4.2 Cost analysis and financial model

Based on the site classification task which initiated by the company Quiet Media, this research contributed to establish a model for machine learning classification to solve this question. However, by means of the launch of some crowdsourcing and micro-tasking project platforms on the Internet, the classification tasks could be solved manually as well. Therefore, in terms of the cost analysis, we plan to carry out estimates cost in a comparative way, which compares the costs of manual labeling with the cost of deployment of our classification model.

4.2.1 Estimate costs input

The structure of estimate costs is organized according to the methodology of project cost management outlined in the book A Guide to the Project Management Body of Knowledge published by PMI in 2017. All theoretical foundations and research methods used in this chapter also derives from this book as an instruction of a fundamental resource.

According to the definition in the book PMBOK Guide, estimate cost is defined as the process of developing an approximation of the cost of resources needed to complete project work (PMBOK Guide, 2017). To organize the process of the cost analysis 3 keys metrics must be defined, which are Inputs, Tools and Techniques and Outputs.

Table 21. Estimate cost Inputs

Estimate costs Task 1 – Define inputs and evaluate inputs techniques			
Inputs	Inputs techniques	Manual Labeling	Classification model based on machine learning
Cost Management Plan	Units of measure	month	month
	Level of precision	Rubles 995.59 to Rubles 1,000	Rubles 995.59 to Rubles 1,000
	Level of accuracy	±10%	±10%
Project Documents	Project schedule	From 2022 to 2026	From 2022 to 2026
	Resource requirements	Yandex Toloka platform	Yandex cloud solutions
Enterprise Environmental Factors	Market conditions	Performers by manually labeling websites or graphs used for machine learning to improve algorithms and receive rewards.	Basically, applying in the improvements of business performance in the advanced companies which rely on big data analysis
	Published commercial information	Standard costs for material and equipment (such as office, software, notebook etc.)	Standard costs for material and equipment (such as office, software, notebook etc.)
	Exchange rate and inflation	Platform is settled in US dollars. Exchange rate basis on April 23,2022 USD/RUB=73.5 (Bank of Russia, April 2022)	Exchange rate basis on April 23,2022 USD/RUB=73.5 (Bank of Russia, April 2022)

Source: [Author Research]

Tools and Techniques in the process of estimate costs were depicted 8 types in the PMBOK Guide distinct from criteria and metrics based on the different projects.

As for this research's estimates costs objectives, the parametric estimating is the best technique because of its higher levels of accuracy depending on the sophistication and underlying data built into the model for both manual labeling and classification model. To control variables, some parameters need to be initiated.

- Estimation of costs separately be calculated by month.
- The number of the employed performer to complete the work by month is 4.
- Salary of employed performer calculated by 160 hours by month, while salary for specialist calculated monthly.
- The number of model specialist is 3, which are respective responsible for different fields (data scientist, model developer and model supporter), as this research is based on 3 person's work.

Besides, the estimation of salary of employed performers refers to the salary of Yandex Toloka platform tasks filtering by category "Classification". In this category the tasks are mainly binary classification tasks, which are analogical to our project task to label images from sites to 2 classes.

Analogical tasks and salaries are summarized in the table below.

Table 22. Estimation of the employed performer's salary

Analogical tasks in the category classification	Salary ($ by hour)
Classification of images	0.058
Verify flavors of products	2.835
Is the answer correct	1.101
Is car parking entrance in the image(s)?	0.28
Detect errors in audios (PSER)	0.90
Argument classification of tweets about COVID-19 health mandates	0.661
Classify advertiser's keyword relevance to search query	1.102
Rate the audio (MOS)	0.078
Which skin types is the product suitable for?	1.519

Check if query is English or not	1.143

Source: [Author Research]

We take the average of these 9 tasks salaries as the baseline of our task's salary, which is $0.96 by hour, by exchange rate USD/RUB=73.5, result in 70.56 rubles by hour.

The estimation of hardware costs refers to the online integration cloud solution platform, we compared 2 platforms which are Yandex Cloud and VK cloud solutions. The services and prices are summarized in the table below.

Table 23. Estimation of the hardware costs

Virtual machine service	Yandex Cloud (Price by month, rubles)	VK cloud solutions (Price by month, rubles)
Virtual Servers for deployment production (4 CPUs, RAM 8GB, SSD 10GB)	4755.9	5470
Virtual Servers for test (2 CPUs, RAM 4GB, SSD 10GB)	2437.5	2800
Virtual Servers for data science modeling (16 CPUs, RAM 32GB, SSD 80GB)	19500	22400

Source: [Author Research]

Based on the scale and complexity of our task, it is rational to rent 3 virtual machine each month to implement separate mission to provide the technical support in the process of the deployment classification model. By comparing the prices on the 2 platforms, Yandex Cloud platform is more cost-effective, we will take its price as our estimation on cost of virtual servers.

It is supposed to employ 3 personnel to provide development and support in the process deployment of the machine learning model, which are data scientist, data developer and administrator by one person. We referred to 3 career portal websites as the benchmark of the jobs' salary by rubles monthly. The salaries are summarized in the table below.

Table 24. Salary references of 3 career portal websites in rubles by month

Career	payscale.com (Salary in rubles)	salaryexplorer.com (Salary in rubles)	Glassdoor.com (Salary in rubles)
Data Scientist	111,717	162,000	176,538
Data Developer	107,908	95,100	98,200
Administrator	51,667	47,600	36,053

Currency: USD/RUB=73.5(Bank of Russia, April 2022)

Source: [Author Research]

The 3 websites mentioned rely on their respective technological software, basically based on the industry, location, and career marketing investigation to give advice on the salary. The advantages of these website information are time-sensitive, comprehensive coverage of occupations and accessible. However, the disadvantages are not authoritative enough, data sources are not clear enough. In summary, the information from the career portal website could be used only for reference.

Moreover, based on the survey from the Russian software association RUSSOFT, by the beginning of 2019 the average salary of experts of Russian software companies run at about ₽100 thousand. In the area 'Development of software, consulting services and other associated services. It amounts to ₽169 thousand including both remuneration for labor and necessary payments to insurance and pension funds (Russian software industry 16-th Annual Survey, 2019).

According to HeadHunter.com, the salary of data scientists is found in the below vacancies, which is from 170,000 rubles to 250,000 rubles.

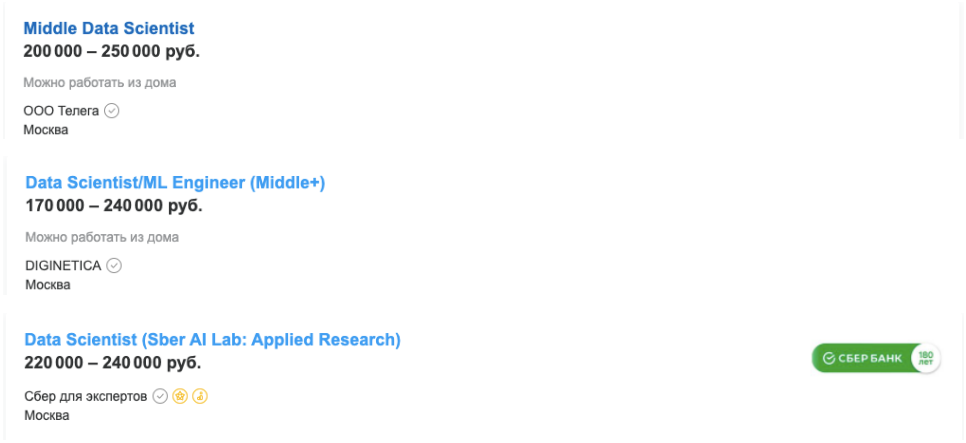


Figure 23. Salaries of data scientist on the website hh.ru

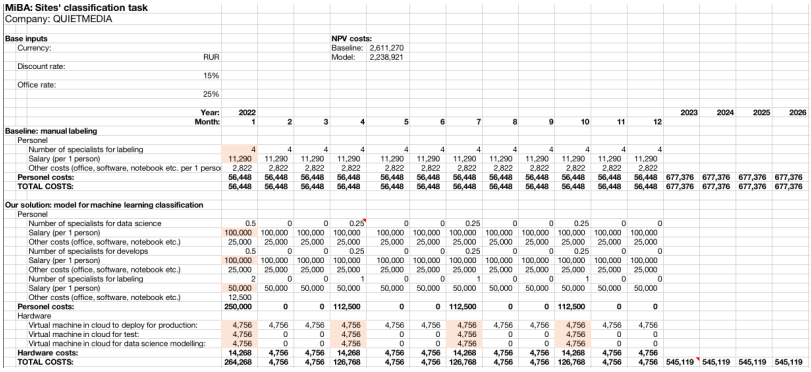
Source: [hh.ru]

Overall, after taking into consideration of 3 sources (career portal website, industry report and job site) of the salary on the data scientist, data developer and administrator. Basically, we could estimate the salary of data scientist as 220,000, the salary of data developer as 95,000, and the salary of administrator as 70,000 rubles by month.

Other costs prepared to use in construction of infrastructure of the mobile devices, basic office supplies, software and emergency fund. The office rate was definite by 25% based on the salary. Discount rate was definite by 15%.

Since the rapid change of websites, new websites are constantly being generated and eliminated. To maintain the good adaptability of the model, we need to re-test and re-train the model with a new training set within 3 months.

4.2.2 Estimate cost Output



MiBA: Sites' classification task																
Company: QUIETMEDIA																
Base inputs				NPV costs:												
Currency:				Baseline:	2,611,270											
RUR				Model:	2,238,921											
Discount rate:																
15%																
Office rate:																
25%																
Year:	2022												2023	2024	2025	2026
Month:	1	2	3	4	5	6	7	8	9	10	11	12				
Baseline: manual labeling																
Personel																
Number of specialists for labeling	4	4	4	4	4	4	4	4	4	4	4	4				
Salary (per 1 person)	11,290	11,290	11,290	11,290	11,290	11,290	11,290	11,290	11,290	11,290	11,290	11,290				
Other costs (office, software, notebook etc. per 1 perso	2,822	2,822	2,822	2,822	2,822	2,822	2,822	2,822	2,822	2,822	2,822	2,822				
Personel costs:	**56,448**	**56,448**	**56,448**	**56,448**	**56,448**	**56,448**	**56,448**	**56,448**	**56,448**	**56,448**	**56,448**	**56,448**	**677,376**	**677,376**	**677,376**	**677,376**
TOTAL COSTS:	**56,448**	**56,448**	**56,448**	**56,448**	**56,448**	**56,448**	**56,448**	**56,448**	**56,448**	**56,448**	**56,448**	**56,448**	**677,376**	**677,376**	**677,376**	**677,376**
Our solution: model for machine learning classification																
Personel																
Number of specialists for data science	0.5	0	0	0.25	0	0	0.25	0	0	0.25	0	0				
Salary (per 1 person)	100,000	100,000	100,000	100,000	100,000	100,000	100,000	100,000	100,000	100,000	100,000	100,000				
Other costs (office, software, notebook etc.)	25,000	25,000	25,000	25,000	25,000	25,000	25,000	25,000	25,000	25,000	25,000	25,000				
Number of specialists for develops	0.5	0	0	0.25	0	0	0.25	0	0	0.25	0	0				
Salary (per 1 person)	100,000	100,000	100,000	100,000	100,000	100,000	100,000	100,000	100,000	100,000	100,000	100,000				
Other costs (office, software, notebook etc.)	25,000	25,000	25,000	25,000	25,000	25,000	25,000	25,000	25,000	25,000	25,000	25,000				
Number of specialists for labeling	2	0	0	1	0	0	1	0	0	1	0	0				
Salary (per 1 person)	50,000	50,000	50,000	50,000	50,000	50,000	50,000	50,000	50,000	50,000	50,000	50,000				
Other costs (office, software, notebook etc.)	12,500															
Personel costs:	**250,000**	**0**	**0**	**112,500**	**0**	**0**	**112,500**	**0**	**0**	**112,500**	**0**	**0**				
Hardware																
Virtual machine in cloud to deploy for production:	4,756	4,756	4,756	4,756	4,756	4,756	4,756	4,756	4,756	4,756	4,756	4,756				
Virtual machine in cloud for test:	4,756	0	0	4,756	0	0	4,756	0	0	4,756	0	0				
Virtual machine in cloud for data science modelling:	4,756	0	0	4,756	0	0	4,756	0	0	4,756	0	0				
Hardware costs:	**14,268**	**4,756**	**4,756**	**14,268**	**4,756**	**4,756**	**14,268**	**4,756**	**4,756**	**14,268**	**4,756**	**4,756**				
TOTAL COSTS:	**264,268**	**4,756**	**4,756**	**126,768**	**4,756**	**4,756**	**126,768**	**4,756**	**4,756**	**126,768**	**4,756**	**4,756**	**545,119**	**545,119**	**545,119**	**545,119**

Figure 24. Financial model on cost estimates (Detail see at **Appendix 2**)

Source: [Author Research]

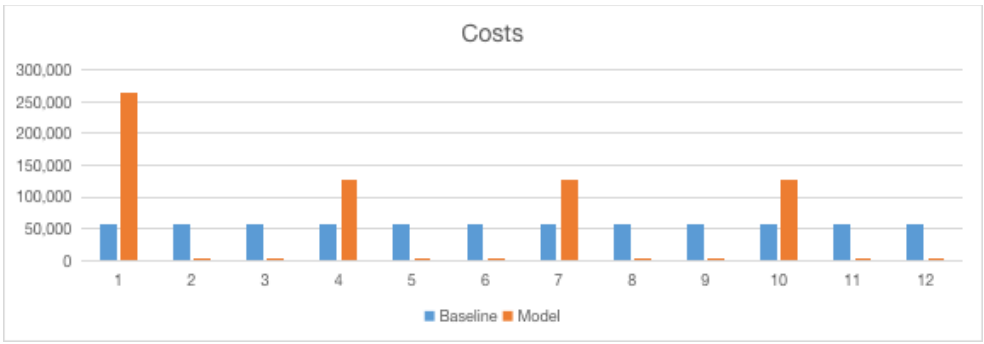


Figure 25. Distribution of the cost estimates by month 2022

Source: [Author Research]

In conclusion, in the duration from 2022 to 2026, estimate cost for task solution by our classification machine learning model is 2,238,921 rubles, however, the estimate cost of the manual labeling as the baseline is 2,611,270 rubles. Generally, although the cost is high from the beginning of the deployment due to the operation and application of the work, it helps to reduce the cost on the further workload, while manual labeling requires continuous output. What's more, NPV are positive both in manual labelling and classification model methods, which means the investment in the task for classification is benefit and profitable for company Quiet Media.

4.3 Risk analysis of model deployment

A machine learning classification model deployment process is very complex. It is affected by many uncertain factors from model deployment initiation to completion and faces many risks.

In the process of software project management, to implement effective risk management activities, it is necessary to identify and evaluate possible risks. This requires the study of effective software project risk assessment methods to provide support for project managers to formulate risk control plans and implement project management activities.

4.3.1 Identify risks

A categorical approach to risk management helps to integrate the activities of the risk analysis. To structure the risk categories, a risk breakdown structure (RBS) is a common way. (See **Figure 26** below)

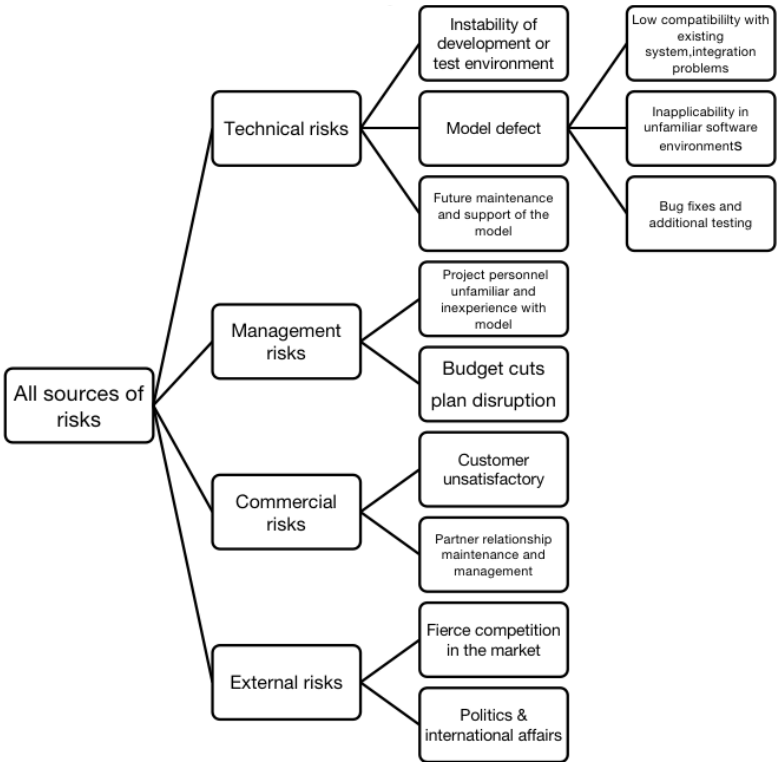


Figure 26. Risk identification structure

Source: [Author Research]

4.3.2 Risk level evaluation and risk responses

Table 25. Risk level evaluation

Risks	Risks Level	Risk Reponses
Instability of development or test environment	Low	Consult the rented virtual server platform for the cause of the problem, debug the equipment, and replace the service

Low compatibility with existing system, integration problems	Medium	Consult experts to find efficient ways to test, adjust compatibility mode to find the optimal solution
Inapplicability in unfamiliar software environments	Low	Consult experts to adjust the compatibility mode and replace the virtual server
Bug fixes and additional testing	Medium	Find optimal alternatives during testing and debugging
Project personnel unfamiliar and inexperience with model	Low	Communicate with model developers about model features, solve application difficulties, and enhance guidance and proficiency
Budget cuts & plan disruption	Low	Confirm the budget amount with senior leadership before deploying the model, and confirm that the application of the model is consistent with the company strategy
Customer unsatisfactory	Low	Investigate the situation, make compensation and prepare for the aftermath
Partner relationship maintenance and management	Low	Strengthening liaison with partners through corporate PR to stabilize relationship and corporation
Fierce competition in the market	Medium	Keen insight into the market situation to adjust the market strategy at any time, conduct consumer analysis, competitor analysis and marketing to maintain competitiveness
Politics & international affairs	Low	Actively pay attention to the international situation and exchange rate changes, and adjust market strategies in a timely manner

Source: [Author Research]

4.4 Business recommendation

The conclusive insights of the above analysis can be summarized through the following points:

- Applying classification model to give solution to commercial and provide improvements already has precedent in the market research, which has approved the classification model solution meets market demand, possesses great market potential and of necessity of deployment to develop business competitivity.

- The cost of applying an integrated platform to improve the business performance is high, which could create huge pressure on the company's budget, not coordinate with the development strategy and economics strategy of company Quiet Media.
- Risk level of deployment classification model has been evaluated from low to medium. Overall, based on our analysis, there will little serve risk in the process of the future business application. The risk exists in a medium-low level, the coordination with partner and clients also makes decisive impact on the whole project.
- The cost of deployment classification model is lower than the cost of manual labeling, moreover, the sustainability of the model could reduce the expenses in the company Quiet Media future development.

To address those issues, and improve profitability and efficiency of Quiet Media program, the following set of recommendations was prepared. (See **Table 26** below)

Table 26. Strategic recommendations summary

Strategic recommendations	Descriptions
Deploy the model for machine learning classification to improve the business performance and optimize services.	Applying advanced solutions based on machine learning technology is the trend of contemporary business, also the optimal way to meet the company's sustainability development strategy. The machine learning classification model established through our research has the advantages of low cost and high accuracy. Deployment model in business could effectively solve problems and create business value for the company Quiet Media.
Develop and strengthen the existing functions of the platform and expand new sections for new services.	QIT platform developed by Quiet Media has several practical functions, but it is not perfect and comprehensive. On the one hand, at the same time of maintaining existing functions and services, it should continue to enhance its functionality. On the other hand, it is important to develop new functions to meet the growing needs of consumers.
Establish customer feedback mechanism	Quiet Media could obtain timely feedback from customers on services through user surveys, market research and other methods to improve business performance. Establishing a good customer feedback mechanism can enable companies to gain insight into consumer behavior in any time and adjust strategic plans timely.

Source: [Author Research]

Conclusion

The research in this thesis builds a high-performance text classification model for Quiet Media to automatically distinguish websites with inappropriate content and evaluates the model through risk analysis and cost analysis to improve the progress of Quiet Media's advertising business. Based on the research of related literature, the survey of market environment, industry research, the application of learned knowledge of machine learning and neural networks, and the understanding of the problem, the company's practical problem is solved constructively.

In the data preparation and model building stages, we introduced the whole steps including data crawling, data preprocessing, text vectorization, data preliminary data analysis and exploration. We used modules like BeautifulSoup and urllib to extract and collect data from websites. Using regular expression to preprocess the data and vectorized text content by implementing TF-IDF. Also fitting curve and wordcloud were used to extract insights from the data and helped us collecting better understanding about the structure of them. In model building stage, cross validation method was used to avoid the potential impact of unbalanced dataset from each category. Multiple metrics such as recall, precision, f1-measure, AUC-ROC were used in order to evaluate the performance of the model to help us choose the best algorithm, parameters and model. Although factors such as the structural quality of the site's content or the quality of the data cannot be assured, we carefully compared multiple approaches at each step and chose the one that best fit our model. In the end, we obtained a model with an AUC-ROC score of 0.989 using the LSTM model. In addition, the pipeline was built to test the behavior of the model and compared the method of manual classification. The result show that, neural networks have better advantages over manual classification and are more suitable for classification of large-scale data, At the same time, it can save a lot of human resources and time resources, thus increasing corporate profits.

In terms of product promotion, we look for a lightweight pornographic website classification tool by using P / MF model, and we focus on this part of the market segment. Through communication with Quiet Media, we further focused on the segment market of pornographic web page screening for advertising. In addition, we also analyzed the cost budget and risk assessment in the product deployment section. By building a financial model for budget evaluation, we can conclude that the cost of the model built using data science-based methods is lower than the cost of manual labeling in the long run. For the risk assessment, based on the basic theory of project management, we classify the risks by type, and in general the risk level of our classification model is low risk.

At the end of the paper, we combine the actual situation of company Quiet Media with the characteristics of our model and provide business recommendations for the company in three aspects, such as technology and platform services.

In the further research, we think the model could be updated in multilingual support and to fit more application scenarios. With larger dataset to be trained, the accuracy would be higher. In addition, most of adult websites would be closed in some day, so the training dataset need to be updated when the tool iterates.

References

Aggarwal, C.C., & Zhai, C. (2012). Mining Text Data. *Springer US*.

Araba, A.M., Memon, Z.A., Alhawat, M., Ali, M., & Milad, A. (2021). Estimation at Completion in Civil Engineering Projects: Review of Regression and Soft Computing Models. *Knowledge-Based Engineering and Sciences*.

Bank of Russia, (2022, April 23) Official exchange rates on selected date https://www.cbr.ru/eng/currency_base/daily/

Brown, J.A., & Wisco, J.J. (2019). The components of the adolescent brain and its unique sensitivity to sexually explicit material. *Journal of adolescence, 72*, 10-13.

Buber, E., & Diri, B. (2019). Web page classification using RNN. *Procedia Computer Science*, *154*, 62-72.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T.P., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide.

Chen, H., Ma, Q., Lin, Z., & Yan, J. (2021). Hierarchy-aware Label Semantics Matching Network for Hierarchical Text Classification. *ACL*.

Chinnadurai, J. (2017). A Framework for Detecting Phishing Websites using EDA algorithm and URL based Website Classification. *International Research Journal of Innovations in Engineering and Technology*, *1*(2), 10.

Dong, K., Guo, L., & Fu, Q. (2014). An adult image detection algorithm based on Bag-of-Visual-Words and text information. *2014 IEEE 10th International Conference on Natural Computation (ICNC)*, 556-560.

Duncan, W.R. (1996). A GUIDE TO THE PROJECT MANAGEMENT BODY OF KNOWLEDGE.

Espinosa-Leal, L., Akusok, A., Lendasse, A., & Björk, K. (2019). Website Classification from Webpage Renders.

Fanaei, M.A., Tahmasbi-Sarvestani, A., Fallah, Y.P., Bansal, G., Valenti, M.C., & Kenney, J.B. (2014). Adaptive content control for communication amongst cooperative automated vehicles. *2014 IEEE 6th International Symposium on Wireless Vehicular Communications (WiVeC 2014)*, 1-7.

Glazkova, A., Egorov, Y., & Glazkov, M. (2020). A Comparative Study of Feature Types for Age-Based Text Classification. *AIST*.

Goodfellow, I.J., Bengio, Y., & Courville, A.C. (2015). Deep Learning. *Nature, 521*, 436-444.

Haddadi, H., Hui, P., Henderson, T., & Brown, I. (2011). Targeted Advertising on the Handset: Privacy and Security Challenges. *Pervasive Advertising*.

Heimerl, F., Lohmann, S., Lange, S., & Ertl, T. (2014). Word cloud explorer: Text analytics based on word clouds. *2014 47th Hawaii international conference on system sciences* (pp. 1833-1842). *IEEE.*

Hernández, I., Rivero, C. R., Ruiz, D., & Corchuelo, R. (2014). CALA: An unsupervised URL-based web page classification system. *Knowledge-Based Systems, 57*, 168-180.

Internet live stats. (2022). Total numbers of websites. https://www.internetlivestats.com/watch/websites/

Iqbal, M., Abid, M. M., Khalid, M. N., & Manzoor, A. (2020). Review of feature selection methods for text classification. *International Journal of Advanced Computer Research*, *10*(49), 2277-7970.

Korobov, M. (2015). Morphological analyzer and generator for Russian and Ukrainian languages. *International conference on analysis of images, social networks and texts* (pp. 320-332). Springer, Cham.

Leiva, F.I., & Kuschel, K. (2020). HMSolution and the search for product-market fit. *Emerald Emerging Markets Case Studies, 10*, 1-20.

Liu, H., He, Y., Hu, Q., Guo, J., & Luo, L. (2020). Risk management system and intelligent decision-making for prefabricated building project under deep learning modified teaching-learning-based optimization. *PLoS ONE, 15*.

Liu, S., & Hao, W. (2021). Forecasting the scheduling issues in engineering project management: Applications of deep learning models. *Future Gener. Comput. Syst., 123*, 85-93.

Maharasi, M., Jeyabharathi, P., & Sivasankari, A. (2013). Text Categorization Using First Appearance And Distribution Of Words.

Meusel, R., Mika, P., & Blanco, R. (2014). Focused crawling for structured data. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (pp. 1039-1048).

Narkhede S. (2018). Understanding AUC - ROC Curve https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Owens, E.W., Behun, R.J., Manning, J.C., & Reid, R.C. (2012). The Impact of Internet Pornography on Adolescents: A Review of the Research. *Sexual Addiction & Compulsivity, 19*, 122 - 99.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Louppe, G., Prettenhofer, P., Weiss, R., Weiss, R.J., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res., 12*, 2825-2830.

Pornhub. (2021). Age distribution of pornhub.com visitors in Russia in 2021 [Graph]. In Statista. Retrieved April 22, 2022, https://www.statista.com/statistics/661813/age-distribution-of-pornhub-visitors-in-russia/

PwC. (2021). Digital advertising expenditure in Russia from 2018 to 2024 (in billion U.S. dollars) [Graph]. In Statista. Retrieved April 21, 2022, from https://www.statista.com/statistics/260681/digital-advertising-spending-in-russia/

Qaiser, S., & Ali, R. (2018). Text mining: use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications*, *181*(1), 25-29.

QuietMedia. n.d. Quiet Media https://qt.media/

Rathore, V. S., & Singh, N.(2019). Plan of Proficient URL Based Web Page Classification Utilizing NLP.

Romney, B. (2020). Screens, Teens, and Porn Scenes: Legislative approaches to protecting youth from exposure to pornography. *Social Science Research Network*.

Rozenberg, N.V., & Korotkova, D.A. (2021). DIGITAL ADVERTISING COMMUNICATION CHANNELS IN RUSSIA AND THE UNITED STATES OF AMERICA.

Russian Association of Communication Agencies. (2021). Distribution of advertising expenditure in Russia from 2000 to 2020, by medium [Graph]. In Statista. Retrieved April 21, 2022, https://www.statista.com/statistics/1025726/russia-advertising-expenditure-share-by-medium/

Shabudin, S., Sani, N. S., Ariffin, K. A. Z., & Aliff, M. (2020). Feature selection for phishing website classification. *Int. J. Adv. Comput. Sci. Appl*, *11*(4), 587-595.

Shahid, F., Zameer, A., & Muneeb, M. (2020). Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. *Chaos, Solitons, and Fractals, 140*, 110212 - 110212.

Shawon, A., Zuhori, S.T., Mahmud, F., & Rahman, M.J. (2018). Website Classification Using Word Based Multiple N -Gram Models and Random Search Oriented Feature Parameters. *2018 21st International Conference of Computer and Information Technology (ICCIT)*, 1-6.

Shi, X., Chen, Z., Wang, H., Yeung, D., Wong, W., & Woo, W. (2015). Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. *NIPS*.

Shubina, A., Palekhova, L., & Shvets, O.D. (2013). PROS AND CONS OF TRADITIONAL AND INTERNET MARKETING.

Tavor, T. (2012). ONLINE ADVERTISING DEVELOPMENT AND THEIR ECONOMIC EFFECTIVENESS. *Australian Journal of Business and Management Research*.

Techopedia. (2018). Mobile Network Operator (MNO) https://www.techopedia.com/definition/27804/mobile-network-operator-mno

Wang, Z., Wang, P., Huang, L., Sun, X., & Wang, H. (2022). Incorporating Hierarchy into Text Encoder: A Contrastive Learning Approach for Hierarchical Text Classification. *ArXiv, abs/2203.03825.*

Yun-tao, Z., Ling, G., & Yong-cheng, W. (2005). An improved TF-IDF approach for text classification. *Journal of Zhejiang University-Science A*, *6*(1), 49-55.

Zhang, X., Zhao, J.J., & LeCun, Y. (2015). Character-level Convolutional Networks for Text Classification. *ArXiv, abs/1509.01626*.

Appendix 1 Code Detail

Table1. Required module in project

```
import numpy as np
import pandas as pd
import boto3
import json
import matplotlib.pyplot as plt
pd.set_option('display.max_columns', None)
import os
import re
from tqdm.auto import tqdm
from random import randint
from urllib.request import (
    Request,
    urlopen,
    URLError,
    HTTPError,
    ProxyHandler,
    build_opener,
    install_opener)
from time import sleep
from bs4 import BeautifulSoup
from langdetect import detect
import socket
import collections
import seaborn as sns

import pymorphy2 as pm
import nltk
import multiprocessing
from multiprocessing import Pool
from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer
from sklearn.metrics import roc_auc_score
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix, precision_score,
recall_score, f1_score
from sklearn_pandas import DataFrameMapper
from sklearn.preprocessing import FunctionTransformer
from sklearn.pipeline import make_pipeline
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import cross_val_score, KFold, GridSearchCV
from sklearn.model_selection import StratifiedKFold, KFold
from sklearn.impute import SimpleImputer
import joblib
from sklearn.preprocessing import LabelEncoder,OneHotEncoder
```

```
46 from keras.layers import LSTM, Activation, Dense, Dropout, Input, Embedding
47 from keras.preprocessing.text import Tokenizer
48 from keras.preprocessing import sequence
49 from keras.callbacks import EarlyStopping
```

Source: [Author Research]

Table 2. Request the site and access the content

```
 1 def get_content(url_page, timeout,proxies=None, file=False):
 2     counts = 0
 3     content = None
 4     MIN_TIME_SLEEP=1
 5     MAX_TIME_SLEEP=3
 6     MAX_COUNTS=2
 7
 8     while counts < MAX_COUNTS:
 9         try:
10             request = Request(url_page)
11               request.add_header('User-Agent', 'Mozilla/5.0 (Windows NT 6.1; WOW64)
12 AppleWebKit/537.36 (KHTML, like Gecko) Chrome/74.0.3729.169 YaBrowser/19.6.1.153
13 Yowser/2.5 Safari/537.36')
14             if proxies:
15                 proxy_support = ProxyHandler(proxies)
16                 opener = build_opener(proxy_support)
17                 install_opener(opener)
18                 context = ssl._create_unverified_context()
19                 response = urlopen(request, context=context, timeout=timeout)
20             else:
21                 response = urlopen(request, timeout=timeout)
22             if file:
23                 content = response.read()
24             else:
25                 try:
26                     content = response.read().decode(response.headers.get_content_charset())
27                 except:
28                     content = None
29             break
30         except URLError as e:
31             counts += 1
32             print('URLError | ', url_page, ' | ', e, ' | counts: ', counts)
33             sleep(randint(counts * MIN_TIME_SLEEP, counts * MAX_TIME_SLEEP))
34         except HTTPError as e:
35             counts += 1
36             print('HTTPError | ', url_page, ' | ', e, ' | counts: ', counts)
37             sleep(randint(counts * MIN_TIME_SLEEP, counts * MAX_TIME_SLEEP))
38         except socket.timeout as e:
39             counts += 1
40             print('socket timeout | ', url_page, ' | ', e, ' | counts: ', counts)
```

```
        sleep(randint(counts * MIN_TIME_SLEEP, counts * MAX_TIME_SLEEP))
    return content
```

Source: [Author Research]

Table3. Extract text content by using BeautifulSoup

```
def get_text(html):
        soup = BeautifulSoup(html, 'html.parser')
        for script in soup(['script', 'style']):
                script.extract()
        page_text = soup.get_text()
        for ch in ['\n', '\t', '\r']:
                page_text = page_text.replace(ch, ' ')
        return ' '.join(page_text.split())
```

Source: [Author Research]

Table 4. Text preprocessing by using pymorphy2

```
def preprocessing(sentence, as_list=False):
    MORPH = pm.MorphAnalyzer()
    nltk.download('stopwords')
    STOPWORDS = nltk.corpus.stopwords.words(LANG)
    s = re.sub('[^а-яА-Яa-zA-Z]+', ' ', sentence).strip().lower()
    s = re.sub('ё', 'е', s)
    function_words = {'INTJ', 'PRCL', 'CONJ', 'PREP'}
    lemmatized_words = list(map(lambda word: MORPH.parse(word)[0], s.split()))
    result = []
    for word in lemmatized_words:
        if word.tag.POS not in function_words:
            result.append(word.normal_form)
    result = [w for w in result if w not in STOPWORDS]
    if as_list:
        return result
    else:
        return ' '.join(result)
```

Source: [Author Research]

Table 5. Text vectorization

```
def get_vector_text(df, col, max_df , min_df ,tfidf=True):
    tfidf_text = []
    for i, row in df.iterrows():
        tfidf_text.append(row[col])
    print(len(tfidf_text))
    if tfidf:
        vectorizer = TfidfVectorizer(
            ngram_range=(1, 1),
```

```
            max_df=max_df,
            min_df=min_df
        )
    else:
        vectorizer = CountVectorizer(
            ngram_range=(1, 1),
            max_df=max_df,
            min_df=min_df
        )
    features = vectorizer.fit_transform(tfidf_text)
    print(features.shape, np.max(features), np.min(features))
    result = pd.concat(
        [
            df,
            pd.DataFrame(
                features.todense(),
                columns=vectorizer.get_feature_names()
            )
        ],
        axis=1
    )
    return result, vectorizer
```

Source: [Author Research]

Table 6. Classsification model pipeline building by using cross-validation

```
def crossval1(x,y,MIN,MAX,cvfolds,classifier):
    scores = np.zeros((cvfolds,1))
    skf = StratifiedKFold(n_splits=cvfolds,random_state=30, shuffle=True)
    cv_j=0
    for train_index ,test_index in skf.split(x,y):
            x_train, x_test = x.iloc[train_index], x.iloc[test_index]
            y_train, y_test = y.iloc[train_index], y.iloc[test_index]
            preprocessor = DataFrameMapper([
            ('proc', TfidfVectorizer(
                min_df=MIN,
            max_df=MAX,stop_words='english'))
            ], input_df=True, df_out=True)


            pipe = make_pipeline(preprocessor,classifier).fit(x_train,y_train)
            scores[cv_j] =roc_auc_score(y_test, pipe.predict(x_test))
            recall = recall_score(y_test, pipe.predict(x_test))
            f1 = f1_score(y_test, pipe.predict(x_test))
            precision = precision_score(y_test, pipe.predict(x_test))
            print(recall,f1,precision)
            joblib.dump(pipe, r'%s.model'%classifier)
            cv_j+=1
```

```
    print(np.mean(scores))
    return scores
```

Source: [Author Research]

Table 7. LSTM model building and training

```
inputs = Input(name='inputs',shape=[max_len])
layer = Embedding(max_words+1,128,input_length=max_len)(inputs)
layer = LSTM(128)(layer)
layer = Dense(128,activation="relu",name="FC1")(layer)
layer = Dropout(0.5)(layer)
layer = Dense(2,activation="softmax",name="FC2")(layer)
model = Model(inputs=inputs,outputs=layer)
model.summary()
model.compile(loss="categorical_crossentropy",optimizer='adam',metrics=["AUC",'Recall','Precision'])
model_fit = model.fit(train_seq_mat,train_y,batch_size=128,epochs=10,
                      validation_data=(val_seq_mat,val_y),
                      callbacks=[EarlyStopping(monitor='val_loss',min_delta=0.0001)])
```

Source: [Author Research]

Appendix 2 Cost Analysis

Table 1. Base input

Currency:	RUR
Discount rate:	15%
Office rate:	25%

Source: [Author Research]

Table 2. NPV costs

Baseline:	2611270
Model:	2271516

Source: [Author Research]

Table 3. Baseline: manual labeling, 2022

	Month:	1	2	3	4	5	6	7	8	9	10	11	12
Personel	Number of specialists for labeling	4	4	4	4	4	4	4	4	4	4	4	4
	Salary (per 1 person)	11290	11290	11290	11290	11290	11290	11290	11290	11290	11290	11290	11290
	Other costs (office, software, notebook etc. per 1 person)	2822	2822	2822	2822	2822	2822	2822	2822	2822	2822	2822	2822
Personel costs:		**56448**	**56448**	**56448**	**56448**	**56448**	**56448**	**56448**	**56448**	**56448**	**56448**	**56448**	**56448**
TOTAL COSTS:		**56448**	**56448**	**56448**	**56448**	**56448**	**56448**	**56448**	**56448**	**56448**	**56448**	**56448**	**56448**

Source: [Author Research]

Table 4. Baseline: manual labeling, 2023 – 2026

Year:	2023	2024	2025	2026
Personel costs:	**677376**	**677376**	**677376**	**677376**
TOTAL COSTS:	**677376**	**677376**	**677376**	**677376**

Source: [Author Research]

Table 5. Our solution: model for machine learning classification, 2022

	Month:	1	2	3	4	5	6	7	8	9	10	11	12
Personel	Number of specialists for data science	1	0	0	0	0	0	0	0	0	0	0	0
	Salary (per 1 person)	160000	160000	160000	160000	160000	160000	160000	160000	160000	160000	160000	160000
	Other costs (office, software, notebook etc.)	40000	40000	40000	40000	40000	40000	40000	40000	40000	40000	40000	40000
	Number of specialists for develops	1	0	0	0	0	0	0	0	0	0	0	0
	Salary (per 1 person)	95000	95000	95000	95000	95000	95000	95000	95000	95000	95000	95000	95000
	Other costs (office, software, notebook etc.)	23750	23750	23750	23750	23750	23750	23750	23750	23750	23750	23750	23750
	Number of specialists for server administration	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1
	Salary (per 1 person)	70000	70000	70000	70000	70000	70000	70000	70000	70000	70000	70000	70000
	Other costs (office, software, notebook etc.)	17500	17500	17500	17500	17500	17500	17500	17500	17500	17500	17500	17500
	Personel costs:	**168125**	**8750**	**8750**	**88438**	**8750**	**8750**	**88438**	**8750**	**8750**	**88438**	**8750**	**8750**
Hardware	Virtual machine in cloud to deploy for production:	4756	4756	4756	4756	4756	4756	4756	4756	4756	4756	4756	4756
	Virtual machine in cloud for test:	2438	0	0	2438	0	0	2438	0	0	2438	0	0
	Virtual machine in cloud for data science modelling:	19500	0	0	19500	0	0	19500	0	0	19500	0	0
	Hardware costs:	**26693**	**4756**	**4756**	**26694**	**4756**	**4756**	**26694**	**4756**	**4756**	**26694**	**4756**	**4756**
	TOTAL COSTS:	**194818**	**13506**	**13506**	**115131**	**13506**	**13506**	**115131**	**13506**	**13506**	**115131**	**13506**	**13506**

Source: [Author Research]

Table 6. Our solution: model for machine learning classification, 2023 – 2026

Year:	2023	2024	2025	2026
TOTAL COSTS:	**568571**	**568571**	**568571**	**568571**

Source: [Author Research]