



An Algorithm for Multi-Domain Website Classification


Mohammad Aman Ullah, International Islamic University, Chittagong, Bangladesh

 <https://orcid.org/0000-0001-6260-5433>

Anika Tahrin, International Islamic University, Chittagong, Bangladesh

 <https://orcid.org/0000-0001-9308-4984>

Sumaiya Marjan, International Islamic University, Chittagong, Bangladesh

 <https://orcid.org/0000-0002-5474-9768>

ABSTRACT

The web is the largest world-wide communication system of computers. The web has local, academic, commercial and government sites. As the types of websites increases in numbers, the cost and accuracy of manual classification became cumbersome and cannot satisfy the increasing internet service demands, thereby automated classification became important for better and more accurate search engine results. Therefore, this research has proposed an algorithm for classifying different websites automatically by using randomly collected textual data from the webpages. This research also contributed ten dictionaries covering different domains and used as training data in the classification process. Finally, the classification was carried out using the proposed and Naïve Bayes algorithms and found the proposed algorithm outperformed on the scale of accuracy by 1.25%. This research suggests that the proposed algorithm could be applied to any number of domains if the related dictionaries are available.

KEYWORDS

Classification, Dictionary, Dynamically, Feature, Matching, Text, Website

INTRODUCTION

A website may be an assortment of web content, images, videos or alternative digital assets that are hosted on one or more internet server sometimes accessible via the net. Websites are frequently devoted to a selected issue, starting from diversion and social networking to providing news and education. With the intensification in the variety of sites, the requirement for website classification gains attraction (Wang et al., 2010). Website classification is a very challenging issue and needs human expertise if it is done manually. The work cost of these standard classifications is also winding up progressively high, and this classification has turned out to be gradually troublesome (Deng, 2012). To overcome the usual classification problem of the websites, many machine learning algorithms such as naive Bayes, support vector machine, random forest, etc. have been used by the researchers in their works. In most of the research works, classic algorithms were used for the classification and classified only single domain.

DOI: 10.4018/IJWLTT.2020100104

This article, originally published under IGI Global's copyright on October 1, 2020 will proceed with publication as an Open Access article starting on January 28, 2021 in the gold Open Access journal, International Journal of Web-Based Learning and Teaching Technologies (converted to gold Open Access January 1, 2021), and will be distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

This research has proposed an algorithm for classifying different websites automatically by using randomly collected textual data from the web pages. This research also contributed ten dictionaries covering different domains and used as training data in the classification process. The classification was done by both the proposed and Naïve Bayes algorithm and found the proposed algorithm outperform the naïve Bayes on the scale of accuracy by 1.25%. This study suggests that the proposed algorithm could be applied to any number of domains provided that the related dictionaries are available.

Therefore, the contributions of this research are:

1. Proposal of an algorithm to classify the websites of different domains such as food, business, education, shopping, travel, and social media, etc.;
2. Creation of different dictionaries to characterize the said domains;
3. Improvement of the accuracy of web search.

This paper is structured as follows: section 2 includes a narrative of related works; section 3 represents the problem Statement. In Section 4, the description of the methodology is provided. Section 5 contains the details of data collection and preprocessing. Section 6 includes description regarding experiments and proposed algorithm. Section 7 is all about experiment results and analysis. The comparison is discussed in section 8. Finally, in section 9 conclusions and future work directions are discussed.

RELATED WORK

Most of the work done so far emphasized the classification using classic classifier and classify at most two to three domains. (Patil et al., 2012) applied a Naïve Bayes algorithm to categorize the websites using the content of the homepages. As per them, web pages could be classified to a more specific category using different feature sets. (Roul et al., 2014) have classified the Web Document using the Association Mining technique. The classification was done using the frequent itemsets created by the Frequent Pattern (FP) Growth algorithm. Final classification was done on the feature set by Naïve Bayes classifier. A simple method was proposed by (Slamet et al., 2018) for web scraping to find the job vacancy from the Search Engine using Naïve Bayes classifier. (Klassen et al., 2010) works on Web document classification by keywords using random forests, their experiment showed that, increasing in domain reduces the accuracy of the classifier.

(Meng et al., 2017) employed an automatic detection system for filtering course website from the general results of search engine. With a motivation to improve classification accuracy, research on Web Classification was conducted by (Abidin et al., 2016) by using an algorithm with the revised n-Grams Word Dictionary. They also examine an algorithm to assess its efficacy in the binary classification problem. A Machine Learning Approach was applied by (Akanbi et al., 2014) for Phishing Website Classification and showed the importance of viable features in phishing website classification. They have used the Decision Tree algorithm along with pruning decision tree to reduce the complexity. For differentiating phishing websites from logical ones based on features, a rule-based intellectual phishing websites classification was done by (Mohammad et al., 2014). They have also applied the same algorithm for predicting phishing websites with an intention to trim down the false-negative rate and thus improved accuracy.

(Bruining, 2015) works on Automatic Classification of Business Websites. This research concerns on Automatic Classification of Business Websites via a supervised learning method. This research was only focused on the Dutch market. It uses custom software for crawling the websites and the Weka-toolkit as the main machine-learning toolkit. This research shows it is possible to scrape websites belonging to businesses and classify them with an accuracy of 70-75%. (Abdessamed et al., 2015) classified the web site on the basis of URL and Content of the websites considering the

case perspective of Algerian Vs. non-Algerian. For making the distinction between the web pages of Algerian and non-Algerian cases, they have trained and used different machine learning algorithm and finally, a novel way was introduced that could combine URL and content information. As per them, for extracting more nation centered contents, this application could be expanded to social networks and different blogging platforms.

PROBLEM STATEMENT

From the study, it is evident that, in most cases, websites are classified on a single domain and thus few dictionaries exist to classify the sites. Also, the classification is done using classic classifier such as naive Bayes, support vector machine, random forest, etc. So, a necessity creates for working on multi-domain website classification; and creation of more dictionaries to fulfill the needs of website classification.

METHODOLOGY

Figure 1 shows the Proposed Model for Multi-domain Website Classification. First of all, text data were collected from the websites. The text was then preprocessed by removing the punctuations, numbers, single letters, and stop words. The preprocessed data was then tokenized into unigram. Finally, this research removed the missing values from those texts. This data was subsequently used as Test Data for the models. For the training data of the model, the dictionaries were created with the help of the internet sources. The proposed and the Naive Bayes classification models were then trained with the data from the dictionaries. This study, then tests the classification of websites using both the models with the same testing data. Finally, this research evaluated the performance of both the models based on the different parameters of confusion matrix such as accuracy, precision, recall, and F-score.

DATA COLLECTION AND PREPROCESSING

Text data were collected with the help of software called Octoparse using the URL of the websites as shown in Figure 2. Data of ten different domains such as Food, Education, Shopping, Traveling, Car and Car reservation, Hotel, Freelancing, Photography, News and Social Media were collected. For each domain, at least ten different URL of the websites were used to retrieve the data. This data was then used as test data after preprocessing. The information for creating the dictionaries for the above domains was collected by exploring different related sites and internet sources. These dictionaries were used as training data after preprocessing. Figure 4 shows one dictionary among ten dictionaries.

In reality, quality raw data are very rare. Therefore, preprocessing of raw data is immensely necessary for bringing out some quality data. This study has done the following pre-processing such as removal of additional value or stop words, removal of punctuations, numbers, single letters, tokenizing of the text, converting the text into lower case, as well as removing the missing values on each of the text documents collected for different domains. The preprocessing thus produces some useful and quality data as shown in Figure 3. The total procedure was also followed for creating the dictionaries.

EXPERIMENT

The implementation starts with the training of the proposed model by the defined dictionaries. Then, preprocessed text for each site is loaded as testing data. These data are matched with each of the ten dictionaries, find and save the paired values of different variables. This research then compared the values of the variables to find the most significant value. From the matched variables of the most significant value, the proposed model identifies the domain. In the case of the Naïve Bayes algorithm,

Figure 1. Proposed model for multi-domain website classification

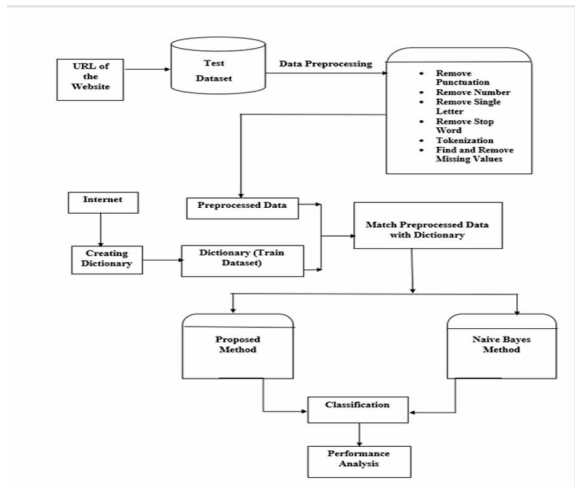


Figure 2. URL's of food website



both the training and testing data were labeled with the corresponding domain name as shown in Figure 5 and Figure 6. The model was then trained with the training data. Finally, the model was tested with the testing data to identify the domain. The performance of both the models was evaluated using different parameters of the confusion matrix such as Accuracy, Precision, Recall, and F-measure as shown in the next section.

PROPOSED ALGORITHM

Step 0: Begin

Step 1: Set dictionaries to train set (manually created unigram dictionaries);

Step 2: Set preprocessed data to test set (text from websites);

Figure 3. Test data before and after preprocessing

```
TED is a nonpartisan nonprofit devoted to spreading ideas,
usually in the form of short, powerful talks.
TED began in 1984 as a conference where Technology, Entertainment
and Design converged, and today covers almost all topics – from
science to business to global issues – in more than 110
languages. Meanwhile, independently run TEDx events help share
ideas in communities around the world.

after preprocessing

[1] "ted"           "nonpartisan"  "nonprofit"    "devoted"
"spreading"       "ideas"        "usually"      "form"
[9] "short"        "powerful"     "talks"        "ted"
"began"           "conference"   "technology"    "entertainment"
[17] "design"       "converged"    "today"        "covers"
"almost"          "topics"       "science"       "business"
[25] "global"      "issues"       "languages"     "meanwhile"
"independently"   "run"          "tedx"          "events"
[33] "help"        "share"        "ideas"
"communities"    "around"       "world"
```

Step 3: Match test data with each dictionary (training data), calculate the term frequency and record in variables v1, ..., v10 (label with the dictionary name)

Step 4: Compare the values from v1, ..., v10 and find the one with maximum match

Step 5: print the variable name (with maximum match) with label and get the classified site

Step 6: End

EXPERIMENT RESULTS AND ANALYSIS

In total, data from 80 websites were provided to the proposed model as the test data, and the model successfully classified 73 of them to the desired domain. The classification result generated by the proposed model is presented in Table 1.

Figure 4. Sample word dictionary of food

food	beer	chinese	noodles	delivered	family	tastes
stir	wine	exotic	cheese	house	visits	pure
shake	grill	fatty	pie	ensures	country	shapes
restauran	steak	digest	chocolate	favourite	company	streets
kitchen	chef	simple	red	freshness	free	things
eatery	store	diet	champagne	well	fruits	machine
cafe	meal	cusine	milk	harvest	appetite	mediterr
bistro	party	unusual	mushrooms	better	parlour	nourishes
bill	pasta	junk	green	place	markets	effort
breakfast	plate	organic	cow	owners	lunchtim	carry
check	platter	vegetarian	farmer	blended	use	launched
cup	lemonade	processed	price	mealtime	colours	craved
dessert	salad	frozen	bakery	hamful	eggs	experien
dinner	main	convenience	flavour	healthful	shaking	grille
dressing	counce	baby	flavours	choice	sorbets	freashest
drink	fries	delicacy	delight	eating	tastes	fresh
fork	fruit	speciality	deliciousness	office	pure	economy
hamburger	ice	suger	world	natural	shapes	desserts
knife	creams	bread	service	animals	streets	fort
lunch	cake	basket	box	working	things	weekend
menu	pizza	soup	ingredients	farmers	machine	events
napkin	delicious	butter	parties	welcome	mediterr	technolog
order	tasty	shrimp	enjoy	seasonal	nourishe	tufnell
buffet	fresh	roast	dine	worked	effort	kerb
salt	healthy	beef	invite	farms	carry	launched
spoon	nutritious	vegetables	cravings	serve	launched	including
water	plain	rice	hand	environment	craved	violet
coffee	spicy	fish	stock	local	experien	recyclable
tea	hot	lasagna	wayne	garden	grille	park
sushi	cold	baked	delivery	weekly	freashes	music
hotel	italian	potato	satisfy	fast	fresh	best
pub	french	spaghetti	catering	family	economy	

Figure 5. Sample of training dataset with a class label for naive Bayes classifier

A	B
word	class
food	food
restauren	food
kitchen	food
travel	hotel
services	hotel
hospitalit	hotel
leisure	hotel
affiliate	market
marketpla	market
credit	market
crunch	market
informer	news
inform	news
airbrush	photo
aperture	photo
black	photo
shopper	shopping
trading	shopping
share	social
messenger	social
traffic	travel
controller	travel
lost	travel
emissions	car
nut	car
basic	education
education	education

Figure 6. Sample of test data set with a class label for naive Bayes classifier

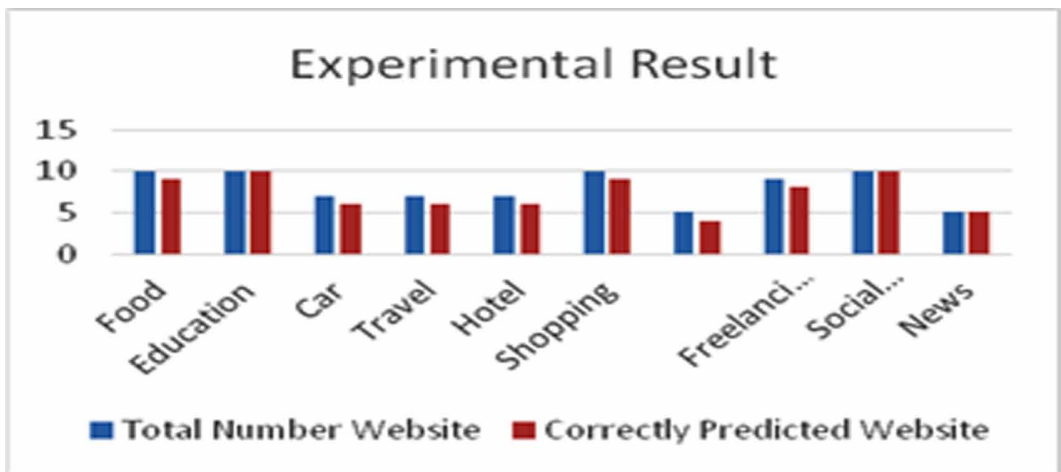
A	B
word	class
envision	social
world	shopping
anyone	news
can	car
transform	social
life	education
accessing	social
world	news
best	photo
learning	education
experience	travel

Figure 7 also shows the classification result between the total number of websites and correctly predicted website. Both the results from Table 1 and Figure 7 clearly show that, the accuracy of the prediction is high.

Table 1. Classification result generated by the proposed model

Domain	Food and Restaurant	Education	car	Travel	Hotel	Shopping	Photography	Freelancing	Social Media	News
Total number of websites	10	10	7	7	7	10	5	9	10	5
Correctly predicted website	9	10	6	6	6	9	4	8	10	5

Figure 7. Classification result analysis



The Accuracy, Precision, Recall, and F-measure of the proposed model are approximately 91%, which also reflects that the prediction rate is high. Detailed are shown below:

Accuracy = (correctly predicted website/total number of websites) *100

Here, Total number of web sites = 80

Correctly predicted website = 73

Number of related documents retrieved

Precision = Total number of documents retrieved =0.91

Number of related documents retrieved

Recall = Total number of documents retrieved. =0.91

(2*Recall*precision)

F-measure = (Recall + Precision)

F-measure = (2*0.91*0.91) / (0.91+0.91) = 0.91

Now, Accuracy = (73/80) *100 = 91.25%

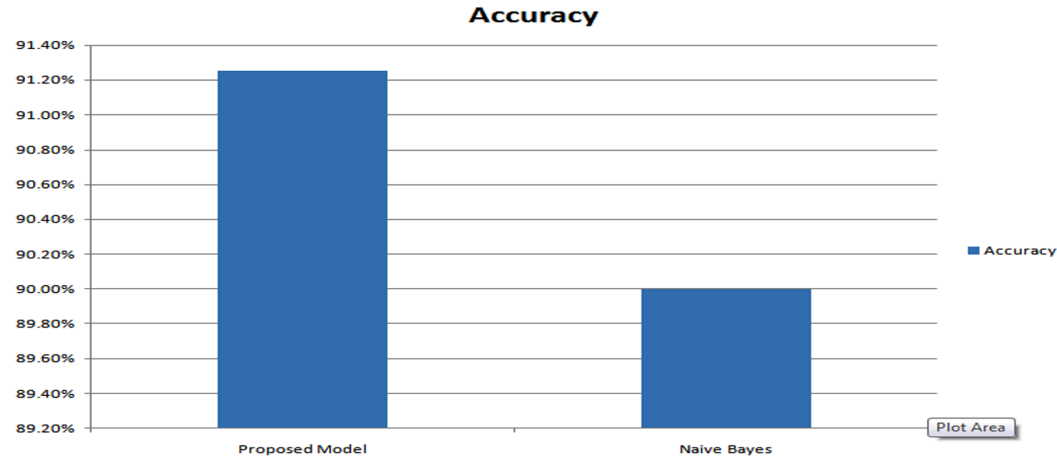
COMPARISONS

In classifying 10 different domains such as Food and Restaurants, Education, Shopping, Tours and Travel, News, Photography, Social Media, Freelancing site, Hotel, Car and Car Reservation system from 80 websites, the proposed model shows the accuracy of 91.25%, whereas the naïve Bayes shows the accuracy of 90%. Hence, the proposed model outperforms the naïve Bayes by 1.25%. Table 2 and Figure 8 show the comparison of results between the proposed model and Naïve Bayes.

Table 2. Comparison of results between the proposed model and Naïve Bayes

Proposed Model	Naïve Bayes
91.25% (Accuracy)	90% (Accuracy)

Figure 8. Comparison of results between the proposed model and Naïve Bayes



CONCLUSION AND FUTURE WORK

With the increase in web sites, the necessity for classification gets attraction. Therefore, an algorithm for website classification is proposed and implemented so that the problem with the existing classification may overcome. In order to meet the classification requirement, ten dictionaries on different domains were created. For simplicity, this research considered unigram dictionaries only. From the experiment, this research concludes that, rich set of dictionaries and appropriate algorithm could well classify the websites. This research also recommends that, the proposed algorithm could be applied to any number of domains. In the Future; this work could be expanded by classifying websites of more domains based on n-grams dictionaries.

REFERENCES

- Abdessamed, O., & Zakaria, E. (2015, April). Web site classification based on URL and content: Algerian vs. non-Algerian case. In *Proceedings of the 2015 12th International Symposium on Programming and Systems (ISPS)* (pp. 1-8). IEEE.
- Abidin, T. F., & Ferdhiana, R. (2016, October). Algorithm for updating n-grams word dictionary for web classification. In *Proceedings of the International Conference on Informatics and Computing (ICIC)* (pp. 432-436). IEEE. doi:10.1109/IAC.2016.7905758
- Akanbi, O., Abunadi, A., & Zainal, A. (2014). Phishing Website Classification: A Machine Learning Approach. *Journal of Information Assurance & Security*, 9(5).
- Bruining, E. (2015). Automatic Classification of Business Websites.
- Deng, F. (2012). Web service matching based on semantic classification.
- Klassen, M., & Paturi, N. (2010, July). Web document classification by keywords using random forests. In *Proceedings of the International Conference on Networked Digital Technologies* (pp. 256-261). Springer. doi:10.1007/978-3-642-14306-9_26
- Meng, R., Zhao, Z., Chi, Y., & He, D. (2017). Automatic Course Website Discovery from Search Engine Results. In *iConference 2017 Proceedings* (Vol. 2). Academic Press.
- Mohammad, R. M., Thabtah, F., & McCluskey, L. (2014). Intelligent rule-based phishing websites classification. *IET Information Security*, 8(3), 153–160. doi:10.1049/iet-ifs.2013.0202
- Patil, A. S., & Pawar, B. V. (2012, March). Automated classification of web sites using Naive Bayesian algorithm. In *Proceedings of the international multiconference of engineers and computer scientists (Vol. 1, pp. 519-523)*. Academic Press.
- Roul, R. K., & Sahay, S. K. (2014). An Effective Approach for Web Document Classification using the Concept of Association Analysis of Data Mining.
- Slamet, C., Andrian, R., Maylawati, D. S. A., Darmalaksana, W., & Ramdhani, M. A. (2018, January). Web Scraping and Naïve Bayes Classification for Job Search Engine. *IOP Conference Series. Materials Science and Engineering*, 288(1), 012038. doi:10.1088/1757-899X/288/1/012038
- Wang, H., Shi, Y., Zhou, X., Zhou, Q., Shao, S., & Bouguettaya, A. (2010, October). Web service classification using support vector machine. In *Proceedings of the 2010 22nd IEEE International Conference on Tools with Artificial Intelligence (ICTAI)* (Vol. 1, pp. 3-6). IEEE. doi:10.1109/ICTAI.2010.9

Mohammad Aman Ullah has received his Bachelor of Science in Computer Science and Engineering (CSE) from the International Islamic University Chittagong in 2008. In 2012, he completed a Master of Business Administration from Premier University Chittagong. He has completed a Master of Science in Computer Science and Engineering from Daffodil International University in 2014. He is at present an academic staff in the Department of Computer Science and Engineering, International Islamic University Chittagong.

Anika Tahrin has received her Bachelor of Science in Computer Science and Engineering (CSE) from International Islamic University Chittagong in 2019. Her research interest is data mining and machine learning. She has successfully completed her undergraduate thesis on website classification.

Sumaiya Marjan has received her Bachelor of Science in Computer Science and Engineering (CSE) from International Islamic University Chittagong in 2019. Her research interest is data mining and machine learning. She has successfully completed her undergraduate thesis on website classification.