Classifying Websites into Non-topical Categories

Chaman Thapa, Osmar Zaiane, Davood Rafiei and Arya Sharma

University of Alberta
{chaman,zaiane,drafiei,amsharm}@ualberta.ca

Abstract. With the large presence of organizations from different sectors of economy on the web, the problem of detecting to which sector a given website belongs to is both important and challenging. In this paper, we study the problem of classifying websites into four non-topical categories: *public*, *private*, *non-profit* and *commercial franchise*. Our work treats each website and all pages from the site as a single entity and classifies the entire website as opposed to a single page or a set of pages. We analyze both the textual features including terms, part-of-speech bigrams and named entities and structural features including the link structure of the site and URL patterns. Our experiments on a large set of websites related to weight loss and obesity control, under a multi-label classification setting using the SVM classifier, reveal that with a careful selection and treatment of features based on keywords, one can achieve an F-measure of 70% and that adding structural, part-of-speech and named entity based features further improves the F-measure to 74%. The improvement is more significant when textual features are not accurate or sufficient.

Keywords: non-topical classification, structural features, topical features, non-topical features, web genre

1 Introduction

The tremendous growth of World Wide Web over the past few years has made it extremely easy for end-users to reach the general mass public by having a web presence. As more people, organizations and governments publish their information on the web, it is important and increasingly difficult to find and filter the desirable information from the web. For example, one may want to know from the website of a health clinic if it is publicly funded hence the treatment expenses are paid by the public health insurance. In such a scenario, associating websites with desirable labels can be helpful in improving the search by linking labels with the search query and allowing the users to filter the websites more easily. Automatic classification of websites can also be helpful in automating the process of creating web directories which takes considerable effort if humans were to label the websites manually.

Website classification can be treated under text classification assuming that a website is a set of web pages or documents. A problem with applying a textual classifier to non-topical classes is that these classes may not be well-described in

text and a richer set of features need to be maintained. The problem is similar to classifying documents based upon the sentiment [3, 4], identifying text genre [2], etc. For example, features such as part-of-speech tags, punctuations, and named entities in addition to words from text have turned out to be useful, as reported in our experiments and also in some past work [5, 9]. Beyond text and part-of-speech patterns, the link structure of the site, URL patterns [12, 11, 10] and HTML tags [8] may also provide additional useful information that can help to correctly classify the website.

In this research, we will be classifying websites into 4 non-topical categories: *public, private, non-profit*, and *commercial franchise*. The non-topical categories that we are concerned with are related to websites that fall under the domain of weight loss/obesity control in Canada. Since many service providers for obesity control have a web presence, classifying the entire website would reveal important facts about these organizations, These facts may inform the users, for example, about the cost and the reliability of a service provided by these organizations. Our domain experts have confirmed that these facts will be useful for obesity patients to efficiently filter the required resources from the web. With the categories that we are concerned about, a website can have more than one label, hence we explore how the independent feature sets based on the link structure of the site, part-of-speech and named entity distribution, and bag-of-words perform in a multi-label classification setting.

In order to classify the entire website, we consider a collection of web pages from a website as a single document. A website can contain hundreds of pages and the inclusion of every page from the website increases the number of features based on bag-of-words. It also takes a significant amount of time to extract the part-of-speech and named entities. More attributes would also mean that the classification takes more time and there is a high probability for the occurrence of noise. In order to mitigate this problem, we analyze dimensionality reduction by selecting the features based on information gain and click-depth of a page, the number of clicks required by the user to reach a page. As the click-depth of the website increases, there is an increase in the number of pages. We analyze how the classifier performs when the word-based features are extracted at click-depth of zero, one and two.

Our contributions include: (1) a study in detecting the business type of an entity from its website, (2) a non-topical website classifier with classes that relate to the business type of the entity a website presents, (3) applying the classifier to the real world domain of weight loss and obesity control in Canada, and (4) an experimental evaluation showing the performance of the classifier and the effectiveness of the features studied.

The rest of the paper is organized as follows. Section 2 reviews the related work and Section 3 presents the dataset and the way it is acquired and the labels are assigned. Section 4 presents the types of features and their selection process and Section 5 reports our results and analysis of the classification. Finally, we draw the conclusions in Section 6.

2 Related Work

Related to our investigation is the body of work on non-topical classification on the Web. Mishne [1] illustrated a supervised classification of blog posts based on the mood of the writer. Turney [3] presented an unsupervised classification of reviews and Pang et al. [4] applied supervised algorithms such as SVM, Naïve Bayes and the maximum entropy to movie reviews. These work on non-topical classification focus on finding the right features that work best for the dataset. Some of the features that have been used for sentiment analysis are unigrams, unigrams combined with bigrams and part-of-speech tags. Bekkerman [5] showed that combining POS-bigrams along with bag-of-words improved the classification accuracy in case of the genre classification.

Dai et al. [6] classified web pages into commercial and non-commercial classes in an attempt to detect the online commercial intent of a page. This work is close to ours as some of the categories overlap, however the authors only used keywords from text and html attributes as features whereas we are combining non-topical features with word-based features. Ester et al. [7] performed a topical classification of websites using a k-order markov model and also treating pages from the same site as a single page. Pierre [8] showed that words from metatags are useful in the classification of websites into industrial categories. His result showed that words from metatags alone can be more effective than words from metatags and html body combined together; however, it also showed that metatags is not widely used by many websites. In our research, the bag-of-words feature comprises of words from metatags, title and the html body. A more recent work by Eickhoff et al. [9] classified web pages based on whether it is targeted towards children or not. They combined both topical and non-topical aspects of a document by using features such as part of speech, shallow texts, html features, and language complexity. They showed that combining topical and non-topical features can work well for non-topical classification.

There is also work on analyzing the structural properties of websites. Amitay et al. [10] used the structure of websites to classify them into eight functional categories and showed that sites with similar functions shared similar link structures. Lindemann and Littig [11] did a thorough study on the relationship between the structure and functionality of the websites. Their work analyzed 1461 websites distributed among five functional categories and reported a strong accuracy using the structural properties. Later, Lindemann and Littig [12] also showed that utilizing both the content and structural properties for website classification performed better than using structural or word based features alone.

From the past work on non-topical classification of documents and websites, it is evident that words are a powerful set of features even for non-topical classification. Results have also shown that combining part-of-speech along with words improves the performance for non-topical classification of documents. Non-topical website classification also benefits from a combined feature set where words are augmented by structural properties. Our work combines and analyzes all three aspects (i.e. the structure of text in the form of part-of-speech tags and named

entities, structural pattern of the website, and bag-of-words) in a real world non-topical website classification task.

3 Dataset Preparation

We used a set of keywords related to weight loss in a search engine to come up with the list of websites. As we were only concerned with organizations providing services related to obesity control/weight loss and having a physical presence in Canada, we built a collection of search queries by appending different city names like Edmonton, Toronto, etc to useful keywords, which were suggested by obesity experts. Some of the keywords used were "obesity clinic", "weight management", "fitness and exercise","diet program" etc. Using these search queries with Google, we came up with a list of websites. We then did an extensive online survey where 77 users participated in labeling the websites. The definition of the categories that were used to label the websites are as follows:

Public: A website providing service that is offered or subsidized by the government.

Private: The service provider has a private firm and is a licensed health care professional or has certification.

Non-profit: A service that has been provided on a non-profit basis.

Commercial Franchise: An organization that provides or sells services or products for profit. In many cases, the organization has many branches (> 2) in different parts of the country and is considered a chain.

We picked only those labels where two or more users agreed upon and where the websites provided service related to obesity control. This was checked through the online survey where the users had the option to tag websites that were not related to obesity control. This helped us filter out many blog sites and web directories. However, obtaining the labels this way did not give us enough labels to populate each category as most of the websites in the search results were either private or franchise. Hence, we also asked one patient and one student to extensively search the web for non-profit and public categories. All the website labels were later verified by an expert and the expert's decision on the label was considered final. The final list of labels we collected comprised of 215 websites where the majority of the labels were private (116). This set was highly imbalanced with more than 50% of the labels comprising of private websites. Since we are measuring the strength of each type of feature, we did not want any bias due to over-fitting in micro and macro measures during evaluation; hence we balanced the dataset by randomly selecting the websites from the over populated category. Table 1 shows the label distribution among various combinations of multi-label categories after the dataset was balanced.

Since the public and non-profit categories had a small number of samples, we kept all the websites in these categories. On the other hand, we under-sampled

Table 1. Website count for various label combination

Label Combination	Number of Websites
Public	25
Non-profit	24
Franchise	21
Private	13
Public,Private	10
Public,Private,Non-profit	2
Private,Franchise	24
Public,Non-profit	6
Total	125

the franchise and private categories such that the final label distribution for each category is: public (43), private (47), franchise (45) and non-profit (32).

In order to capture the features based on the language model, we crawled the websites at various click-depths. The landing page or the homepage of a website is considered at a click-depth of zero. All the links present in the homepage are then considered at click-depth of one and so on. We crawled the internal links of each website at click-depth of zero, one and two. We saved only those files for which the server response was valid and the header had text/html as the content-type. The maximum number of files we crawled for each website was limited to 1000 pages. Table 2 shows the number of pages crawled at each click-depth. Click depth of 2 contains all the pages at depth zero, one and two inclusively. We will use this convention throughout the paper.

Table 2. Number of pages crawled at each click-depth

Depth	Number of Pages	Avg. Page Size (In KB)
Click-depth 0	125	24.9
Click-depth 1	5322	99.05
Click-depth 2	34981	127.91

The structural properties of the websites were captured by crawling the internal links of each website with valid HTML server response up to a depth of ten. The HTML pages themselves were not saved but the URLs pertaining to external and internal links at each depth were recorded to extract the structural properties. We only expanded the internal links and marked the external links as a new external link or previously appeared external link. We crawled an internal link only once. If an internal link appears more than once, we mark them as already visited.

4 Features used for Classification

4.1 Features based on words

Words provide useful cues as to which category a particular website belongs. Analyzing the websites with human eye, we can see that websites in franchise categories mostly contain words such as *order*, *pay*, *success*, and *testimonials*. Private websites mostly contain terms like *doctor*, *clinic*, *physician* and other common medical terms. Public websites mostly contain the keyword *government* and non-profit ones often have words like *donate*, *voluntary*. Some of the non-profit organization list themselves as being a not-for-profit organization in their about page. There are many variations of the keyword non-profit expressed as "not for profit", "non profit" or even "not-for-profit". In order to capture this notion, we combined these variations as a single entity: *nonprofit*, using a regular expression.

We followed the bag-of-words approach and extracted word unigrams from HTML documents. We used a HTML parser[1] to extract the text from the body and title of the document. Words from meta-keywords and meta-description tags were then added to the list of unigrams. We then extracted the word-stem for each unigram and represented the word stem in a feature vector using the TF-IDF metric. Since we were dealing with a collection of web pages within a website, we followed a slightly different definition of TF-IDF to normalize the term frequency within a website.

$$\text{tfidf(t,W)} = \frac{\text{tf}}{\text{P}} \times \log\left(\frac{\text{N}}{\text{DF}}\right) \tag{1}$$

In equation 1, *tfidf(t,W)* gives the TF-IDF measure of a term t for the website W. *tf* is the term frequency of the word t i.e. the number of times t occurs in W. P is total number of web pages in W where term t occurs. DF is the document frequency of t with respect to the websites i.e. the number of websites in the dataset in which the term t occurred. N represents the total number of websites in the dataset. We discarded any term whose document frequency (DF) is less than three.

4.2 Part-of-speech and Named Entity based features

Part-of-speech provides useful information about the structure of a sentence which can be helpful in capturing the notion of the categories. It has been shown to be useful for text classification where sentences are well formed. However, capturing POS tags from HTML documents can be a bit tricky as HTML documents can mostly contain words as opposed to sentences. In order to extract POS tags from HTML documents, we processed the document to extract groups of text which contain a sentence boundary, (i.e. the symbols "." , "?" and "!"). We only extracted those sentences which contained more than two words and removed

[1] http://lxml.de/

the anchor tags <a> along with any formatting tags <b>, <i> from the sentences. We used NLTK's default tagger to tag the sentences with the simplified tagset [2] and extracted POS-bigrams from each sentence. We used the frequency of each POS-bigram as a feature.

Besides POS tags, several text patterns can help to identify the category of a website. Franchises often indicate price of an item that is being sold in their website. We used a regular expression to extract the patterns of price that is indicated in dollar amount and used the frequency of price patterns in the feature set.

In many cases, franchises also have more than two branches. In order to capture this notion, we extracted the postal addresses from the HTML page. We looked for those URLs that are present in the home page and contain the word stem "about", "contact", "locat", and "map" in the anchor text. We then used a geo extractor [3] that uses a regular expression based method [13] to extract the postal addresses. We used the total number of unique addresses and the number of different provinces as features based on address.

Some organizations often repeat their name many times in their website which can provide useful hint about the category of the website. We extracted organization names from sentences using NLTK's named entity tagger and counted the occurrence of each organization name. We then picked the organization name with highest frequency and used its frequency and occurrence per page as a feature. We also added the total number of unique organization names and the number of organization names per page to the feature set.

4.3 Link Structure and URL-based features

Amitay et al. [10] and Lindemann et al [11, 12] have analyzed many structural features that are useful to classify the websites into functional categories. We used 8 URL features and 8 link structure based features from [10–12]. URL features were extracted from a collection of URLs obtained during the depth-crawl of the website. The URL features included average number of digits in the path, number of sub-domains encountered during the crawling of website depth, average path length, average number of slashes in the path, fraction of PDF/PS, HTML and script files, and number of unique file types obtained by analyzing the file extension from the URL.

Link structures are mainly based on the external (a link pointing to a page outside the website) and internal links (a link pointing to any other page within the website) and the depth at which these links were found. The structural features included from [10–12] were average external depth, average internal depth, maximum depth of the website, average depth, total number of unique URLs, fraction of links at the densest depth, average size of the crawled pages in KB, and fraction of the files having javascript in it.

[2] http://www.nltk.org/
[3] http://www.folkarts.ca/geo/

We added more features which were mostly based on click-depth level. We counted the external links, internal links, and outdegree at each depth and calculated the maximum internal links, maximum external links and maximum outdegree occurring at any click-depth. We also computed average external links per depth, average internal links per depth and average outdegree per depth. Furthermore, we created four bins for each depth d indicating the counts of external, internal, repeated internal, and repeated external links at that depth. During the crawling phase, we noticed that some of the private websites had very few internal links at a shallow depth. On the other hand, some of the public websites contained many internal links at a shallow depth and the size of the website rapidly grew along with the depth. By assigning count bins at each depth we intended to capture this property. We also created three bins for the maximum depth of the websites indicating the depth value between 0-5, 6-10 and > 10. We assigned a value of zero or one to each bin depending upon whether or not the maximum depth of the websites falls under the range.

In Canada, many government websites have the domain name of the format *.gov* and *.gc.ca*. In order to capture the essence of top-level domain (TLD) names, we created binary features based on the various patterns of top-level domain names present in the dataset. Our dataset consisted of 7 different patterns of domain names including *.org*, *.com*, *.ca*, *.net*, *.gov*, *.province.ca* (where *.province* can be any Canadian province in its short form) and *.gc.ca*. We combined *.gc.ca*, *.gov* as a single pattern representing government and used a total of six binary features based on the occurrence of the various TLD patterns.

5 Experiments

We used a SVM based one-vs-all BR method [20] to perform multi-label classification. As is the case with one-vs-all approach, we built 4 binary classifiers, one for each class label, and assigned a class label to the website if the prediction of the classifier is positive. All the experiments were performed in a stratified 10-fold cross-validation setting. We created the folds only once and used the same set of folds throughout the experiment for consistency.

We used the RBF (Radial Basis Function) as the kernel for SVM and performed a grid search to select the best values for the parameters γ and C. We searched for the best values of γ and C by maximizing the F-measure in a 5-fold cross validation setting. One of the training folds was used for grid search to obtain the value of γ and C, and the same value of γ and C was used for the rest of the folds in the 10-fold cross-validation setting. LibSVM [14] was used for grid search and SVM based classification.

5.1 Feature Selection

Feature selection is an important step in classification and helps to improve the performance by selecting the most informative features. We performed two steps of feature selection based on document frequency and information gain. First, we

discarded any features having document frequency less than 3. We then followed the ALA (All Labeled Alignment) [15] based document transformation approach and calculated the Information Gain (IG) to select the features by ranking. For a website with multiple labels, the ALA approach requires having duplicate rows of the website each indicating one of the labels in the set of feature vectors used for classification. Throughout this research, we measured the information gain only from the training fold in the 10-fold cross validation setting. Features were ranked by information gain and the top x% of the features were selected for classification, where the value of x ranges from 10 to 100. The information gain was calculated using the implementation of Weka [19].

We also analyzed how the number of features at each click-depth of the website affects the classification performance. Table 3 shows the number of features per click-depth. We can see that the number of features based on bag-of-words (BOW) greatly increases along with the click-depth. On the other hand, numbers of POS-bigrams are not much affected. As we deal with structural features separately, the number of features based on structure remains the same.

Table 3. Number of features extracted at each click-depth

Feature	**click-depth 0**	**click-depth 1**	**click-depth 2**
BOW	1221	7876	31923
POS-Bigram	248	307	320
Structure	90	90	90

Figure 1 shows the performance of bag-of-words at each depth and at various thresholds of information gain. We can see that bag-of-words has the least performance at depth 0. Upon analyzing the websites, we found that the index page of the website seldom contains important words which are able to classify the website into the non-topical categories. Lack of words like “non-profit” which mostly occur in the “About” page at a click depth of one supports this fact. At depth 1, the highest micro f-measure attained by bag-of-words is 0.65, at an IG threshold of 70%. At depth 2, it performs best with a micro f-measure of 0.70 at an IG threshold of 60%. Although depth 2 has the higher micro f-measure, we notice that at 60% IG threshold, bag-of-words require 19153 features at depth 2 whereas depth 1 utilizes 5513 features. This clearly suggests the addition of noise at depth 2. Due to the large number of features, the training time of the classifier is also more than that of depth 1.

5.2 Classification by various feature types

We deal with three types of features: bag-of-words, POS-bigram and named entity distribution, and structural properties which provide three different views of the website. Bag-of-words represent the explicit information provided by the website, POS-bigram and named entity distribution capture the patterns in text

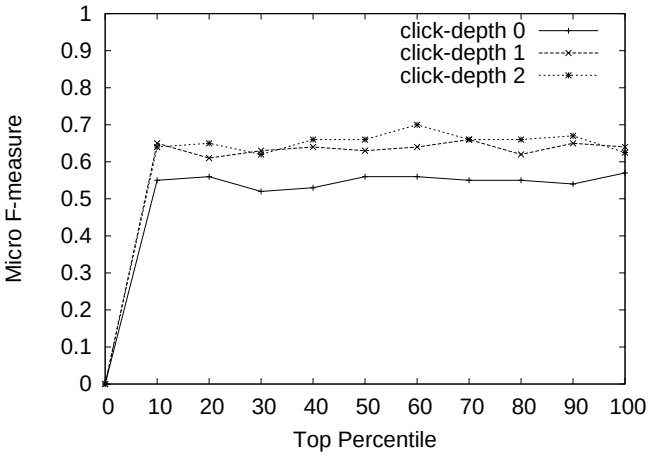

Fig. 1. Micro F-measure for bag-of-words at each click-depth. X-axis denotes top x% of features when ranked by Information Gain where x is between 10 and 100.

and link structure and URL properties give important information about the structure of the website. For each of these views, we created separate feature vector to classify the websites. Table 4 shows the performance of these features at a depth of zero, one and two. As structural properties were crawled separately, they are not related to the number of pages crawled at each depth and has been reported separately without the depth information. Table 4 also shows the information gain threshold at which each set of features perform at its best. Along with the best performance of each set of features, we also include bag-of-words with all the features (at an IG threshold of 100%) to compare the performance gained by selecting important features by IG ranking. Bag-of-words with a threshold on IG seemed to collect the most informative features and as a result performed well out of the three sets of features; however we should note that it also has the largest number of features compared to our non-bag-of-word features. The best performance of POS bigram and named entities require the use of all of their feature at a depth of zero and one, and 90% of their feature at depth two. At depth one and two, the performance of POS bigram and named entities is comparable to that of structural features. Using only the top 30% of their features, structural features attained a micro and macro f-measure of 0.55.

Table 5 shows some of the top features along with their information gain. We could not provide the full list of features due to the space constraints; nonetheless the list gives the informative strength of each set of features. The list corroborates the fact that explicit words on the website provide the most informative features. Words like “voluntary”, “donate”, “nonprofit” are good indicators of a website belonging to the non-profit category, whereas “testimonials”, “llc”, “inc” are good indicators for private and franchise. Keywords “ministry”, “government” help to identify websites from the public category. Bag-of-words has

Table 4. Precision, Recall along with Micro and Macro F-measures at each click-depth

			Public		Private		Non-Profit		Franchise		F-measure	
Depth	Features	IG	P	R	P	R	P	R	P	R	Micro	Macro
zero	BOW	100%	0.67	0.53	0.68	0.48	0.46	0.37	0.71	0.62	0.57	0.56
	POS_Bigram_NE	100%	0.64	0.41	0.56	0.34	0.33	0.18	0.62	0.37	0.42	0.41
one	BOW	70%	0.67	0.53	0.73	0.57	0.69	0.5	0.86	0.71	0.65	0.65
	BOW	100%	0.66	0.46	0.75	0.61	0.63	0.37	0.89	0.75	0.64	0.62
	POS_Bigram_NE	100%	0.67	0.53	0.51	0.59	0.42	0.37	0.54	0.57	0.53	0.52
two	BOW	60%	0.66	0.65	0.63	0.75	0.72	0.56	0.73	0.86	0.70	0.69
	BOW	100%	0.76	0.53	0.62	0.63	0.56	0.28	0.82	0.71	0.62	0.59
	POS_Bigram_NE	90%	0.58	0.55	0.55	0.69	0.45	0.28	0.53	0.46	0.53	0.50
-	Structrue	30%	0.56	0.51	0.54	0.67	0.57	0.59	0.55	0.42	0.55	0.55

the highest value of information gain which is also the reason behind its good performance in classification. Table 5 also shows that the features related to POS bigram are more informative than the patterns captured from named entities. It also confirms that structural features comprising of internal and external links and related statistics at a depth level are good measures for classification. Public websites generally have many pages at a shallow depth, while most private websites have only a few pages at shallow depths. Also, the maximum depth of a website is a good indicator of a government website (at large depth) and a small private clinic (at small depth).

Table 5. Features ranked by Information Gain

Bag-of-words		**POS Bigram + NE**		**Structure**	
volunt	0.318	[ADV PRO]	0.048	Fraction of PDF/PS files	0.155
committe	0.224	[DET ADV]	0.046	Max. external links per level	0.068
collabora	0.222	[TO P]	0.043	Repeated internal links at depth 5	0.061
...		...		...	
ministri	0.203	[VG ADJ]	0.032	Tot. internal links at depth 5	0.059
fund	0.193	[VG PRO]	0.027	Repeated internal links at depth 6	0.056
donor	0.186	[EX ADJ]	0.023	Max.depth between 0 to 5	0.056
donat	0.150	Price Count	0.023	Tot. external links at depth 6	0.055
govern	0.147	[N NP]	0.022	Avg. digit in domain path	0.049
...		...		...	
nonprofit	0.137	[VG WH]	0.018	Number of file types	0.043
community-bas	0.121	Province Count	0.015	Tot. external links at depth 4	0.039
llc	0.046	[ADJ NP]	0.014	Tot. internal links at depth 3	0.032
testimoni	0.036	Address count	0.010	Avg. outdegree per level	0.031
inc	0.023	Org. name count	0.010	Domain with .gov extension	0.011
...		...		...	

5.3 Combining Classifiers

As three different sets of features provide different perspective about the website, we try to capture the notion of all three views by combining the individual classifiers built on each view. Kittler et al. [16] show several ways to combine classifiers. We measured the probabilistic output from SVM using Platt's method [17, 18] and used the sum rule to combine the classifiers based on the three feature types. There are 3 classifier predictions (one from each view) for each website sample. For every sample, we summed up the confidence measure of the classifier based on each feature type for every positive prediction. We then only picked the positive predictions where the sum of confidence measure was more than the best performing threshold of 0.65, which was obtained by trying different threshold values in the range of 0.1 to 1.0 at a step of 0.1. Algorithm 1 shows each step of the process.

Alg. 1 Algorithm used to combine the classifiers in a multi-label setting.

```
1   for each label, l do
2       for each website,w, in test do
3           sum[w] := 0
4       for each independent feature set,f do
5           build a classifier,C, using f,l
6               for each website,w, in test do
7                   [prediction,confidence] = Classify(w,C)
8                   if prediction == 1.0
9                       add confidence to sum[w]
10      for each website,w, in test do
11          if sum[w] > 0.65
12              assign label l to w
```

Table 6 shows the classification result of the combined classifier based on bag-of-words, POS tags and named entities, and structural properties. While combining these classifiers, we choose the IG threshold for each view of feature based on the highest performance of the classifier in terms of micro f-measure as shown in Table 4. For instance, at depth of three we choose classifiers with IG threshold 60%, 90% and 30% for bag-of-words, POS bigrams+Named Entity, and structural property respectively and combine them with the sum rule. Table 6 shows that the combined classifiers performed significantly better than bag-of-words alone at depths of zero and one. The results also show that the performance of the combined classifier at depth 1 is greater than that of bag-of-words at depth 2 and similar to the combined classifier at depth 2. This tells us that collecting words from deeper depth of a website may not be of much help. Moreover, the classifier at depth 1 requires significantly less time to train because the number of features based on bag-of-words is much less than that of

depth 2. The performance of the combined classifier at depth 0, although better than bag-of-words alone at depth 0, is the least of all three depths. This indicates that important words available at some shallow depth greater than zero play an important role in the classification. This also shows that words are important features for non-topical classification and augmenting it with structural properties and POS bigrams and named entities improves the classification.

Table 6. Result of combining the classifiers using sum-rule at a threshold of 0.65.

Depth	Features	Micro F	Macro F
zero	BOW (IG=100%)	0.57	0.56
	Combined	**0.66**	**0.65**
one	BOW (IG=100%)	0.64	0.62
	BOW (IG=70%)	0.65	0.65
	Combined	**0.73**	**0.73**
two	BOW (IG=100%)	0.62	0.59
	BOW (IG=60%)	0.70	0.69
	Combined	**0.73**	**0.74**

6 Conclusion

We showed that structural, part-of-speech and named entity based features help to improve the classification performance when combined with the bag-of-words approach. We analyzed performance of each type of feature at various depths of the website. As the number of words increased with the depth of the website, there was a slight increase in performance of bag-of-words from depth 1 to depth 2, at a cost of significant training and crawling time. We achieved similar performance using words from depth of 1, which takes significantly less amount of time for crawling and training the classifier, combined with structural and part-of-speech based features. The performance of bag-of-words increased significantly from depth of 0 to depth 1 showing that index pages at depth zero seldom contain informative word based features to categorize *public*, *private*, *non-profit* and *franchise*. The non-topical features comprising of part-of-speech, named entity and structural features boosted the performance of bag-of-words at all three depths of zero, one and two. However, the classification performance was the least at depth 0, which showed that words are important features even for non-topical classification. We selected features using the ALA based information gain and combined the most powerful classifiers of each feature-type by sum-rule to achieve 8% gain on micro and macro f-measure at depth of 1.

As topical (bag-of-words) and non-topical (POS bigram and named entity, structural property) features provide different views for the website, combining both of them is extremely helpful. Future work for non-topical website classification might involve trying various ways to combine the classifiers based on these independent views such that performance can be further improved.

References

1. Mishne, G.: Experiments with mood classification in blog posts. In Proc. of the 1st Workshop on Stylistic Analysis Of Text For Information Access, at SIGIR (2005)
2. Kessler, B., Nunberg, G., Schütze, H.: Automatic detection of text genre. In Proc. of the Association of Computational Linguistics Conference, pp. 32-38. (1997)
3. Turney, P. D.: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proc. of the Association of Computational Linguistics Conference, pp. 417-424. (2002)
4. Pang,B., Lee, L.,Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. Computing Research Repository, cs. CL/0205070 (2002)
5. Bekkerman, R., Eguchi, K., and Allan, J.: Unsupervised non-topical classification of documents. Technical Report IR-472, UMass Amherst (2006)
6. Dai, H.K., Zhao L., Nie, Z., Wen, J., Wang, L.,Li, Y.: Detecting Online Commercial Intention (OCI). In Proc. of the World Wide Web Conference, pp. 829-837 (2006)
7. Ester, M., Kriegel, H.,Schubert, M.: Web Site Mining: A New Way to Spot Competitors, Customers and Suppliers in the World Wide Web, In Proc. of the Knowledge and Data Discovery Conference. (2002)
8. Pierre, J. M.: On the automated classification of Web sites. Linköping Electronic Articles in Computer and Information Science 6. (2001)
9. Eickhoff, C., Serdyukov,P., Vries, A.P.: A combined topical/non-topical approach to identifying web sites for children. In Proc. of WSDM Conference, pp. 505-514.(2011)
10. Amitay, E., Carmel, D., Darlow, A.,Lempel, R., Soffer, A.: The Connectivity Sonar: Detecting Site Functionality by Structural Patterns, In Proc. of Hypertext and Hypermedia Conference. Nottingham, United Kingdom. (2003)
11. Lindemann, C.,Littig, L.: Coarse-grained classification of web sites by their structural properties. In Proc. of the 8th International Workshop on Web Information and Data Management. Arlington, VA. (2006)
12. Lindemann, C., and L. Littig. Classification of web sites at super-genre level. In Genres on the web: Computational models and empirical studies. Text, Speech and Language Technology. Dordrecht: Springer. (2010)
13. Yu, Z.: High Accuracy Postal Address Extraction from Web Pages. Master's Thesis. Dalhousie University. Halifax, Nova Scotia, Canada. (2007)
14. Chang,C.-C. and Lin., C.-J.: LIBSVM : a library for support vector machines. ACM Trans. on Intelligent Systems and Technology, 2:27:1-27:27. (2011)
15. Chen, W., Yan J., Zhang, B. , Chen Z., Yang, Q.:Document transformation for multi-label feature selection in text categorization, In Proc. of Seventh IEEE International Conference on Data Mining, pp. 451-456, USA. (2007)
16. Kittler, J.,Hatef, M., Duin, R. P. W., Matas, J.: On Combining Classifiers. IEEE Trans. on Pattern Analysis and Machine Intelligence.Vol. 20, pp. 226-239. (1998)
17. Platt, J.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers. (2000)
18. Lin, H.-T.,Lin, C.-J.,and Weng, R. C.: A Note on Platt's Probabilistic Outputs for Support Vector Machines. Machine Learning, 68(3), 267-276. (2007)
19. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update; SIGKDD Explorations, 11(1). (2009)
20. Tsoumakas, G. and Katakis, K.: Multi label classification: An overview. International Journal of Data Warehousing and Mining, 3(3). (2007)