Master Thesis in Statistics and Data Mining

Functionality classification filter for websites

Lotta Järvstråt

Abstract

The objective of this thesis is to evaluate different models and methods for website classification. The websites are classified based on their functionality, in this case specifically whether they are forums, news sites or blogs. The analysis aims at solving a search engine problem, which means that it is interesting to know from which categories in a information search the results come.

The data consists of two datasets, extracted from the web in January and April 2013. Together these data sets consist of approximately 40.000 observations, with each observation being the extracted text from the website. Approximately 7.000 new word variables were subsequently created from this text, as were variables based on Latent Dirichlet Allocation. One variable (the number of links) was created using the HTML-code for the web site.

These data sets are used both in multinomial logistic regression with Lasso regularization, and to create a Naive Bayes classifier. The best classifier for the data material studied was achieved when using Lasso for all variables with multinomial logistic regression to reduce the number of variables. The accuracy of this model is 99.70 %.

When time dependency of the models is considered, using the first data to make the model and the second data for testing, the accuracy, however, is only 90.74 %. This indicates that the data is time dependent and that websites topics change over time.

Acknowledgment

I would like to express my deepest appreciation to all those who helped me succeed with this thesis.

I would like to give special thanks to all the staff at Twingly, especially Magnus Hörberg, for providing the data and this interesting problem. It has been a very pleasant experience working with you.

I would also like to thank my supervisor Professor Mattias Villani, who has helped and guided me through the wonderful world of text mining. Your advice and support meant a lot to me.

Furthermore I would also like to thank my opponent Emma Lee-Bergström for her improvement suggestions and discussions about the thesis. Thank you Emma, they were really good comments.

Last I would like to thank my loved ones who have supported me in different ways during this thesis work, both with encouraging me and with opinions about my work. Thank you all.

Contents

List of Figures

List of Tables

1 Introduction

This chapter provides the background information, the aim of the thesis and a summary of previous work in this area.

1.1 Background

A lot of research has been done classifying websites by topics, for example [1] and [2]. To classify websites based on their functionality, however, is not as common, albeit quite as important. Functional classification is to classify a website based on purpose, in the present study whether the website is used as a forum, blog or news site. Web crawling, which is extracting information from websites for use by search engines where the web site purpose may be of interest for the search result, is one example where this could be important. The data in this thesis is extracted from a web search engine, which in this case was a blog search. A blog search establishes its results from pinging, where pinging is a service for those who want their blogs to be searchable in different search engines. A ping is a push mechanism by which a blog notifies a server that its content has been updated. This gives the search engines an easy way of knowing when a blog has been updated and it is therefore able to continuously provide updated results for searches. The ping service, however, causes a problem: the ability to ping a site is not unique for blogs; other websites can ping as well, which may lead to search results from other types of websites as well. It is therefore necessary to classify websites based on their functionality.

Text mining is needed to analyze the content of websites. Another name for text mining is "text analytics", a way of making qualitative or "unstructured" data into variables usable by computer. Qualitative data is descriptive data that cannot easily be measured in numbers and often includes qualities such as colour,

texture and textual description. Text mining is a growing area of research due to the massive number of text information provided in electronic documents, both on the web and in other places (such as patient journals, bug reports etc.). Heretofore, texts were gathered manually and were therefore tedious to analyze and compare. Today, despite the rapid growth in available data, the use of high-performance computers and modern data-mining methods allows an ever-increasing value-gain from automated text classification.

The main challenge for text mining is that the data is unstructured and therefore harder to analyze. The number of words in a text can be very large and the data is subsequently often high-dimensional and sparse. There are also difficulties with, for example, noisy data (such as spelling mistakes or text speak), punctuation and word ambiguity. Natural languages are also hard to analyze due to the linguistic structures, the order of the words could be of importance to the analysis [3].

1.2 Aim

The aim of this master thesis is to evaluate a classification method that classifies the functionality of websites with high accuracy. The classification method should be able to recognize blogs, news sites and forums, based on the content of the site.

Another aim is to compare the classification performance of selected models. The selected models are Naive Bayes, which is a common model for text classification, and multinomial logistic regression with Lasso. Multinomial logistic regression was chosen because it is a very simple model, both easy to interpret and computationally fast. Multinomial logistic regression was compared with both features from the word count and HTML code and more advanced features extracted from Latent Dirichlet Allocation (LDA) models for unsupervised topic learning.

Since much of the material on blogs, forums and news sites tend to focus on current topics, there is a risk that classifiers trained

during a certain time period will translate poorly to future time periods when attention has shifted to other topics. The models were therefore also analyzed to assess the influence of time on the classification accuracy of the model.

1.3 Definitions

The following basic definitions are needed to fully understand the problem.

Blog

A blog is a portmanteau of the term web log and is a website for discussion or information. It consists of discrete entries (or blog posts) typically displayed in reverse chronological order.

Forum

An Internet forum, or message board, is an online discussion site where people can discuss in the form of posted messages.

News site

A news site is a site on the web that presents news, often an online newspaper.

HTML

Stands for Hypertext Markup Language, and is a programming language for creating web pages.

Web crawler

A web crawler is a software application that systematically browses the internet for information, sometimes for the purpose of web indexing.

1.4 Related work

Web page classification is much more difficult than pure text classification due to the problem of extracting text and other structural important features embedded in the HTML of the page. In the field of web page classification there are several different areas involved,

one of them is content classification, which means that the web page is classified according to the content of the site, e.g. sport, news, arts etc. This area is probably the most researched area in this field and a lot of methods have been evaluated, for example Naive Bayesian classifiers [1,4], support vector machines (SVM) [2,5], extended hidden Markov models [6], Kernel Perceptron [1], k-nearest neighbour (kNN) [7], different summarization methods [8] and classification by using features from linking neighbours (CLN) [9].

The area for this report is functional classification of websites, which means that the web page is classified according to the function of the site, e.g. forums, news sites, blogs, personal web pages etc. Lindemann et al. [10] used a naive Bayesian classifier to classify web pages based on their functionality. Another study was made by Elgersma et al. [11] and deals with classifying a web page as either blog or non-blog. In that study a lot of models were evaluated, for example SVM, naive Bayesian classifier, Bayesian networks etc.

The information embedded in the HTML can be exploited for classification in many different ways, the most common being to weigh the words depending on which part of the HTML they come from. Another approach is to use the text from the HTML with usual text mining methods and then extract other information such as the outgoing links and relations to other websites.

2 Data

The data was extracted with the help of the company Twingly. The websites were manually classified into the different categories. The raw data entries are the site's url (which is an ID variable), the HTML-coding of the site and the content of the site. The content is the text that is extracted from the website.

Table 2.1: The different data sets

Data set	Extracted	Blogs	News sites (Domains)	Forums (Domains)
1	January 2013	3,543	10,900 (6)	6,969 (4)
2	April 2013	11,600	3,399 (17)	3,400 (17)
Mixed	January and April 2013	15,143	14,299 (17)	10,369 (17)

Table 2.1 shows the different datasets used in this thesis.

The first dataset consists of 3,543 different blogs, 10,900 news pages from six different news sites (Dagens Nyheter, Svenska Dagbladet, Aftonbladet, Expressen, Göteborgs-Posten, Corren) and 6,969 forum pages from four different forum sites (Familjeliv, Flashback, Hembio, Sweclockers). The content was extracted over one week in late January 2013.

Some text data is dependent on the time of publication, especially for the news sites, which can give an over-optimistic view of the models' generalization performance on websites at a later date. For example the events that were top news in January may not be relevant at a later time. To examine if this time dependency matters to the classification filter, another dataset was extracted in April 2013. Using extracted data from only the one time period when the dataset was extracted, would have caused problems, as will be shown later.

The second dataset consists of 3,399 news pages from 17 different news sites (Dagens Nyheter, Svenska Dagbladet, Aftonbladet, Ex-

pressen, Göteborgs-Posten, Folkbladet, Dagens Handel, Eskilstuna-Kuriren, Hufvudstadsbladet, VästerviksTidningen, NyTeknik, Norrbottens-Kuriren, Katrineholms-Kuriren, Sydsvenskan, Dagens Industri, Norrköpings Tidningar, Kristianstadsbladet), 3,400 forum pages from 17 different forum sites (Sjalbarn, Flashback, Fotosidan, Pokerforum, Passagen debatt, Bukefalos, Ungdomar.se, Allt om TV, Garaget, Fuska.se, Sweclockers, Zatzy, MinHembio, AutoPower, webForum, Sporthoj.com, Familjeliv) and 11,660 different blogs.

Both datasets were taken from Swedish websites, which means that most of the texts are in Swedish, although some of the websites are written in other languages, typically English.

2.1 Data cleaning

One big problem with having many observations (web pages) from the same websites is that the start and the end of the texts may be similar due to the underlying structure of the website. For example, the footer of the news sites almost always contains information about the responsible publisher (for example his/her name). This information will of course appear as a strong classifying feature for the news sites with this specific publisher in the training data, but will most likely be useless in classifying other news sites. Therefore the starting and the ending texts of each news sites and forums were removed to eliminate this problem. There are problems with this kind of cleaning however, because the phrase "responsible publisher", which would also appear in the footer, could be an excellent word feature for other news sites. This sort of cleaning was, however, only used for the data material from January 2013, where this problem was largest due to the small number of domains compared to a much larger number of websites. The data from blogs are much more heterogeneous making them easier to use unaltered.

The data cleaning for both data sets also includes converting the text to lower case, removing punctuation and special characters, removing numbers and removing "stopwords" (for example "it", "is"

and “this”). Stemming was rejected, because this did not work well on Swedish words, see Section 5.1. All data cleaning was made using the R-package ’tm’.

2.2 Extracting HTML information

HTML is the layout language for creating a website and it contains a lot of information including the text of the website. What makes it different from normal text is that it has information about where on the site the text appears, such as in the header, information about pictures and links of the sites and sometimes metainformation about the website, for example author and language. In this thesis, apart from the text only the number of links on the HTML was used, though it would be possible to include more advanced features based on the HTML information.

The variable that was extracted from the HTML-code, was the number of links (numlinks), and is shown in detail below.

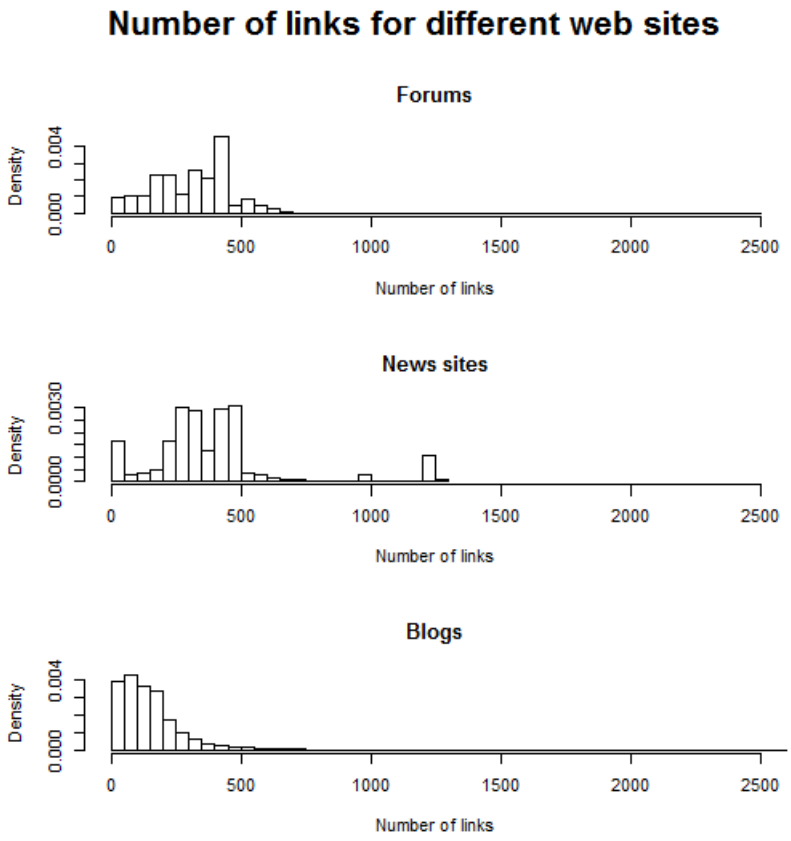

Figure 2.1: Histogram for number of links for the different types of websites

Figure 2.2 shows histograms for the variable numlinks for the different websites categories in the dataset. It can be seen that blogs usually have 250 links or below (the mean is 160), but there are a few outliers with more than 2000 links. Both news sites and forums have typically somewhere between 250 and 500 links (the mean for news sites is 385 and for forums it is 303). The variance for the number of links is highest for news and lowest for forums.

The extraction of the HTML-variable is made in Python with the package `BeautifulSoup`, which can extract information from HTML-code. The text variables were created and modelled in R with different packages. For the variable creation described in Section 3.1 the package *tm* was used, for the multinomial logistic re-

gression the package *glmnet* was used and for the LDA the package *topicmodels* was used.

3 Methods

In this section the different methods that are used in this thesis are explained.

3.1 Variable creation

All text variables are defined as the word count for a specific word, for example the word "test". If "test" appears three times in the text of a website then the value of that variable will be three for that website. This means that there will be a large number of variables in the resulting data set. To reduce them without losing too much information, the words that appeared in less than 0.01% of the websites were removed from the data set. The words that appear in very few websites are probably not good generalized classifiers, and they make the calculations too computationally heavy. This results in 6,908 word variables, using the merged dataset (the same variables were used in the dataset Jan13 and Apr13).

The reason why non-binary features are used instead of using just binary features is that some information about the data would then be lost. This may be of importance when a word appear more than once in a text.

In Figure 3.1 it can be seen that most of the word variables appears in very few web sites. This means that the data is very sparse and the mean is 4.35% and the median is 2.31%. This means that half of the words only appears in 2.31% or less of the websites.

3.2 Multinomial logistic regression

Multinomial logistic regression [12, 13] is used when the response variable can attain many discrete unordered values, often called *categories*. Logistic regression is generally used for data where the response variable only allows two values, but since the case addressed

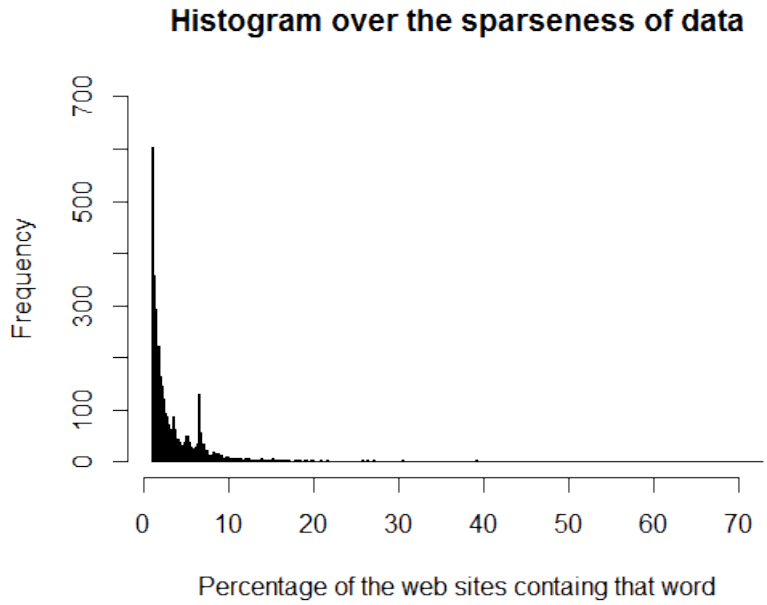

Figure 3.1: Histogram over the sparseness of data

in this thesis can have three values (blog, news site or forum) multinomial logistic regression is used. The logistic regression model comes from wanting to model a transformation of the probabilities of the K categories via linear functions in the covariates, x. The traditional way of modeling this is in terms of $K-1$ log-odds or logit transformations as follows

$$\begin{aligned}
\log \frac{Pr(G=1|X=x)}{Pr(G=K|X=x)} &= \beta_{01} + \beta_1^T x \\
\log \frac{Pr(G=2|X=x)}{Pr(G=K|X=x)} &= \beta_{02} + \beta_2^T x \\
&\vdots \\
\log \frac{Pr(G=K-1|X=x)}{Pr(G=K|X=x)} &= \beta_{0(K-1)} + \beta_{K-1}^T x,
\end{aligned} \tag{3.1}$$

where in this thesis $K=3$. The reason why there is one less equation than there are categories is because the probabilities add up to one,

and one of the categories is the reference category.

In the case when the multinomial logistic regression is used with a penalty method such as Lasso (Section 3.3), it is possible to use a more symmetric approach, where

$$Pr(G = l|X = x) = \frac{e^{\beta_{0l}+x^T\beta_l}}{\sum_{k=1}^{K} e^{\beta_{0k}+x^T\beta_k}}, \tag{3.2}$$

without any explicit reference category. The problem with this approach is that this parameterization is not estimable without constraints since the solution is not unique. Any set of values for the parameters $\{\beta_{0l}, \beta_l\}_1^K$ and the parameters $\{\beta_{0l} - c_0, \beta_l - c\}_1^K$ would give identical probabilities (c is a p-vector). With regularization this problem is naturally solved, because although the likelihood-part of this is insensitive to (c_0, c), the penalty is not.

To find the best model estimate (and not only a unique model) the regularized maximum (multinomial) likelihood is used. The log-likelihood is maximized with respect to β,

$$\max_{\{\beta_{0l},\beta_l\}_1^K \epsilon \mathbb{R}^{K(p+1)}} \left[\frac{1}{N} \sum_{i=1}^{N} logP(G = g_i|X = x_i) \right]. \tag{3.3}$$

3.3 Lasso regularization

When performing a multinomial logistic regression with the new variables described in Section 3.1, there are too many variables to be used directly in the regression. Therefore a variable selection has to be done and Lasso [14] was chosen rather than for example the ridge regression because it not only shrinks the coefficients, but the coefficients are allowed to be exactly zero, thereby also providing a variable selection. Since so many variables were extracted from the dataset to fully characterize the different websites and it is of minor interest which variables are chosen as long as they give a high model prediction accuracy, Lasso was considered the best variable

selection method for this problem.

Lasso employs a penalty function added to the objective function:

$$\lambda * \sum_{j=1}^{p} |\beta_j|,$$

where β_j are the coefficients for the multinomial logistic regression described in section 3.2 and λ is the penalty parameter.

The penalty parameter λ is chosen by K-fold cross-validation. K-fold cross-validation means that the data is split into K equally large subsets. Then the regression is fitted on $K-1$ of the K parts as training set and the Kth remaining part as test set. This is repeated until each of the parts has been used exactly one time as a test set. Then the average error across all K trials are computed and the λ for which the average error is minimal, is chosen.

When the penalty term is added, the function to maximize in multinomial logistic regression is

$$\max_{\{\beta_{0l},\beta_l\}_1^K \epsilon \mathbb{R}^{K(p+1)}} \left[\frac{1}{N} \sum_{i=1}^{N} logP(G = g_i|X = x_i) - \lambda \sum_{l=1}^{K} |\beta_l| \right]. \quad (3.4)$$

3.4 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a topic model introduced in 2001 [15]. "Topic models are algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents." [16] The basic idea is that documents (in this case websites texts) are represented as random mixtures over latent topics, where each topic is characterized by a distribution over the set of used words. To explain the model the following terms are defined:

- A *word* is defined to be an item from a vocabulary indexed by $\{1, \ldots, V\}$. The words are represented by using unit-basis vectors having one component equal to one and all the others equal to zero. Thus, if superscripts are used to denote com-

ponents, the vth word in the vocabulary is represented by a V-vector w such that $w^v = 1$ and $w^u = 0$ for $u \neq v$

- A *document* is a sequence of N words denoted by $\mathbf{w} = (w_1, w_2, \ldots, w_N)$, where w_n is the nth word in the document.
- A *corpus* is a sequence of M documents denoted by $D = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\}$

LDA assumes that each document is generated by the following process:

1. Choose number of words N, where $N \sim Poisson(\xi)$
2. Choose the vector of topics proportions θ, where $\theta \sim Dir(\alpha)$
3. For each of the N words w_n:
   (a) Choose a topic $z_n \sim Multinomial(\theta)$
   (b) Choose a word w_n from $p(w_n|z_n, \boldsymbol{\beta})$, a multinomial distribution conditioned on the topic z_n

The number of topics k, is considered known and fixed, the dimensionality of the variable z. The word probabilities, i.e. the probability of a word being chosen given the topic, are parameterized by a $k \times V$ matrix where V is the size of the vocabulary. This matrix is denoted by $\boldsymbol{\beta}$ where $\boldsymbol{\beta}_{ij} = p(w^j = 1|z^i = 1)$.

The posterior distribution of the hidden topics proportions θ and the topic assigments z given a document is

$$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w}|, \alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)}.$$

To normalize the distribution the function is marginalized over the hidden variables and written in terms of the model parameters:

$$p(\mathbf{w}|\alpha, \beta) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma\left(\alpha_i\right)} \int \left(\prod_{i=1}^{k} \theta_i^{\alpha_i - 1}\right) \left(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} \left(\theta_i \beta_{ij}\right)^{w_n^j}\right) d\theta$$

This function is intractable because the coupling between θ and β in the summation over the latent topics renders the whole posterior distribution mathematically intractable. One way to overcome this problem is to approximate the posterior distribution by a simpler function $q(\theta, \mathbf{z}|\gamma, \phi)$ which depends on the so called variational parameters γ and ϕ. The variational parameters are chosen to minimize the Kullback-Leibler divergence between the true distribution and the variational approximation. This problem can be reformulated as a problem of maximizing a lower bound of marginal likelihood $p(\mathbf{w}|\alpha, \beta)$. To simplify the maximization problem it is assumed that the variational approximation factorizes as follows

$$q(\theta, \mathbf{z}|\gamma, \phi) = q(\theta|\gamma) \prod_{n=1}^{N} q(z_n|\phi_n),$$

where the Dirichlet parameterγ and the multinomial parameters $(\phi_1, \ldots, \phi_N)$ are the free variational parameters. The optimization problem for the variational parameters then becomes:

$$(\gamma^*, \phi^*) = \arg\min_{(\gamma,\phi)} D(q(\theta, \mathbf{z}|\gamma, \phi) ||p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)),$$

where $D(q, p)$ is the Kullback-Leibler divergence between the two densities p and q.

So for each document the variational parameters are optimized and these are then used to find the approximate posterior. Then the model parameters α and β that maximizes the (marginal) log likelihood of the data:

$\ell(\alpha, \beta) = \sum_{d=1}^{M} log\, p(\mathbf{w}_d|\alpha, \beta)$, are estimated by maximizing the resulting lower bound on the likelihood. Then the α and β are used to optimize γ and ϕ again. These two steps are repeated until the lower bound on the log likelihood converges. This algorithm is called the Variational Expectation-Maximization algorithm, or the VEM-algorithm.

3.5 Naive Bayes Classifier

The Naive Bayes classifier [17] is a very common text classifier because of its simplicity. It is based on Bayes theorem, which in this case states that for a document d and a class c

$$P\left(c|d\right) = \frac{P(d|c) * P\left(c\right)}{P\left(d\right)},$$

but since the term $P(d)$ appears in all of the calculations for the probability of a class, this term can be excluded without any loss. This gives the following expression

$$P\left(c|d\right) \propto P(d|c) * P\left(c\right),$$

where $P(c)$ (which is called the prior) is estimated by the relative frequency of that class in the dataset.$P(d|c)$ (or the likelihood function) could also be written as $P(x_{1,}\ldots,x_n|c)$ where $\{x_{1,}\ldots,x_n\}$are the words in a document. This is not possible to calculate because of the large number of variables, which makes an even higher number of combinations. If the assumption that the words are independent of each other is made, then the most likely class can instead be written as

$$c_{MAP} = \arg\max_{c_j \epsilon C} P(c_j) \prod_{i=1}^{n} P(x_i|c_j).$$

This classifier also assumes that the order of the words in the document is random and does not hold any information. This is called the bag-of-words assumption, because the words are treated as a collective as a bag of words without any order.

4 Results

This section presents the results achieved in this thesis. First, the Naive Bayes model fit is presented, then the multinomial logistic regression with Lasso. The observed time dependency is described and finally the use of LDA-topics as features is examined.

4.1 Naive Bayes classifier

Naive Bayes is one of the most common methods used in text classification and a classification is therefore made with both datasets merged including the wordcount variables and the numlinks HTML-variable.

Table 4.1: Contingency table of classification for Naive Bayes with the merged data

	Real label			
Classified as	Blog	News site	Forum	Total
Blog	7,521	876	1,801	10,198
News site	33	5,743	1	5,777
Forum	88	147	2,593	2,828
Total	7,642	6,766	4,395	18,803

In table 4.1 it can be seen that this model does not perform well. The accuracy rate for this model is 84.33%, which is considerably lower than rates reported in previous studies (for example Lindemann et al. [10] achieved an accuracy of 92% in their Naive Bayes approach to classify websites based on their functionality) and thus not considered satisfactory.

4.2 Multinomial logistic regression with lasso

To evaluate the time dependency, only the dataset from January 2013 was used for initial regression. The dataset from April 2013

will then be incorporated to examine the time dependency further.

Each dataset was randomly divided into two equally large datasets, one for training and one for testing. Using the multinomial logistic regression with Lasso regularization 99.62% of all websites were classified correctly. Table 4.2 shows the distribution of the classified observations. As can be seen from this table, the method quite accurately distinguishes between forums and news sites; only one observation is misclassified. It was found that blogs are more difficult to accurately separate from the other categories, probably because of their higher variation in topics.

Table 4.2: Contingency table of classification for multinomial logistic regression

	Real label			
Classified as	Blog	News site	Forum	Total
Blog	1,715	26	39	1,780
News site	3	4,842	0	4,845
Forum	4	1	2,771	2,776
Total	1,722	4,869	2,810	9,401

Appendix A lists the variables that are important for the classification of the websites. Note that the variable extracted from the HTML-code "numlinks" is one of the significant variables. It can be seen that Lasso was able to significantly reduce the number of variables in the model from 6,909 to 257. If all variables had been retained in the model, the model would have been overfitted to the training data, which would lead to a worse fit for the test set.

The presence of variables containing numbers is not fitting, because this is often due to dependencies between some of the websites (for example the ones from Flashback). The number could, for example, be a count of how many threads there were on a forum the day the data material was collected. The numbers in the material are therefore removed and the result for a new multinomial logistic regression with Lasso shown in Table 4.3. This table shows that 99.61% of the websites are classified correctly, which is only slightly

less than when numbers were not removed.

Table 4.3: Contingency table of classification for multinomial logistic regression with the number variables removed

	Real label			
Classified as	Blog	News site	Forum	Total
Blog	1,678	10	6	1,694
News site	13	4,778	2	4,793
Forum	4	1	2,656	2,661
Total	1,695	4,789	2,664	9,148

Appendix B lists the variables chosen by Lasso for each class for the model with the dataset from January 2013. The variables that are bold are negatively correlated with the class; for example, the presence of the curse word "jävla" reduces the probability that the page is a news site. Here, approximately 263 variables are chosen by the Lasso, which is slightly more than when the number variables were in the model. Some of the word variables in this model are directly correlated with some of the domains, for example the word variable "nyhetergpse" which is a subdomain of the news site "Göteborgsposten". Some of the variables seem to be related to the date when the sites were extracted, for example, words like "rysk" (russian) and "snabbmatskedjorna" (the fast-food companies) may be words that are related to the date of extraction. Some of the words, though, seem to be reasonable classifiers like "användarnamn" (username) for forums, "copyright" for news sites and "blogga" (to blog) for blogs.

In the appendix it can also be seen that there are more variables for classifying blogs than for the other categories. This probably depends on the large heterogenity of the blogs.

4.3 Time dependence

After performing the first multinomial logistic regression, a need was identified to include data extracted at different times and with a

larger number of domains, especially forums and blogs. Therefore, a second data material from April 2013 will be used. The result using that material (see section 2) will be presented in this section. The same model as used in section 4.2 was employed with the entire dataset from April 2013 as test data. This gives an accuracy rate of 90.74% (see table 4.4), which is considerably less than the accuracy rate of 99.61% from the previous regression. This means either that the data is time dependent, or that the number of domains was too low in the January 2013 dataset causing the model to overfit that data.

Table 4.4: Contingency table of classification for multinomial logistic regression with the data from April 2013 as test data

	Real label			
Classified as	Blog	News site	Forum	Total
Blog	11,576	466	366	12,408
News site	54	2,850	710	3,614
Forum	31	82	2,324	2,437
Total	11,661	3,398	3,400	18,459

To make the classification filter less time-dependent and less domain-dependent, the two datasets were merged and a new filter was made. This new filter gives an accuracy rate of 99.64%. The contingency table for the test data for this new classification filter is shown in table 4.5.

Table 4.5: Contingency table of classification for multinomial logistic regression with the merged data

	Real label			
Classified as	Blog	News site	Forum	Total
Blog	7,625	30	16	7,671
News site	6	6,734	1	6,741
Forum	12	2	4,379	4,393
Total	7,643	6,766	4,396	18,805

The new variables are shown in Appendix C. The variables with

negative correlation are in bold. It can be seen that some variables are the same as when using the previous filter, but some are new. Some variables come from the HTML-part of the websites, since the web crawler in some cases was unable to extract only text. Some variables such as "Göteborgsposten", "lhc" (a local hockey team) and "vädertrafikhär" (which is a part of a menu on a website) appear to be strongly domain specific, which means that the classification filter still suffers from domain dependency.

4.4 Latent Dirichlet Allocation

In this section LDA will be used to reduce the number of variables in the dataset to a smaller number of topics. The problem with too many variables is that the model tends to become overfitted to the training data. Reducing the variables can rectify this problem. Another benefit with using LDA is that the model can be easier to interpret if the created topics turn out to be intuitively reasonable. LDA is unsupervised, which means that the topics are found in the dataset without considering the classes. This means that the topics do not necessarily have anything to do with the classes.

Here the LDA-topics will be used in a multinomial logistic regression with Lasso where topics correlating with the classes are chosen. A multinomial logistic regression with Lasso using both the word variables, the LDA-topics and the HTML-variable was then fitted to investigate if the accuracy improves with the LDA-topics.

4.4.1 Number of topics

In LDA the number of topics in the model must be decided by the user. Three scenarios are considered with 5, 10 and 20 topics created by the LDA. After this a multinomial logistic regression is fitted using the LDA topics and the HTML-variable numlinks.

Table 4.6: Comparison of LDA with different number of topics

Number of topics	Accuracy rate	Variables used in the model (including numlinks)
5	85.46%	6
10	95.67%	11
20	97.62%	21

As can be seen in table 4.6 these models work well considering that they use a dramatically smaller covariate set (6, 11 and 21 variables) compared to the original data set with 6,906 word variables. All of these models contain numlinks in the final model and even though Lasso is used, all of the variables are retained in the final model. The model with highest accuracy rate is the one with 20 LDA-variables; this model is elaborated further in the following section.

4.4.2 Topics of the LDA

Given the unsupervised nature of LDA, it can be somewhat difficult to find patterns in its inferred topics . In this case, the probabilities for the different topics given the class will be shown as will the ten most common words in each topic.

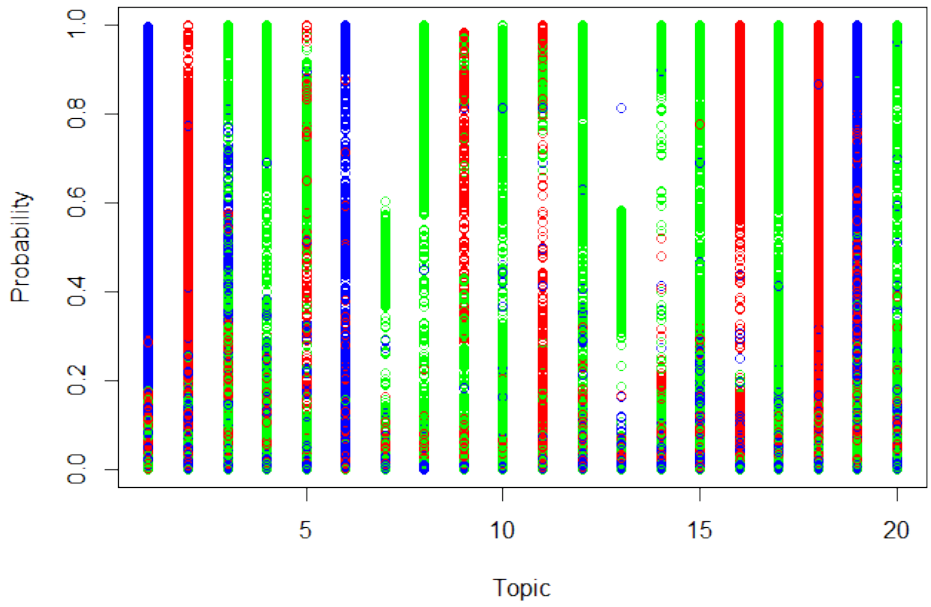

Figure 4.1: The different web sites probability of belonging to a certain topic.

Figure 4.1[1] illustrates the probability that a website belongs to a certain topic. Blogs are blue, news sites green and forums are represented by red. It can easily be seen that topics 1, 6 and 19 are likely to be observed when the website is a blog (blue). Topic 1 contains words like "comments", "April", "March" and "o'clock", topic 6 contains small useful words that are more likely to occur in small talk rather than more serious documents, and topic 19 contains English words. Forums (red) seem to be more likely to belong to topics 2, 16 and 18. Topic 2 contains small talk words, topic 16 contains typical forum words like "registered", "member", "quote", "report" and topic 18 also contains typical forum words like "post", "show", "member" and "forum". News sites seem to be more

[1]This figure should preferably be seen in color. This can be done on the website for this thesis; the web address can be found on the last page.

likely to contain a larger numberof topics, which may be due to the heterogeneity of the sites.

Table 4.7: The ten most common words in topic 1-5

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Forum words	Small talk words	?	?	?
snygga	visa	kommentarer	svenska	paring
kommentarer	endast	mer	mera	är
tweet	säger	ska	ska	för
pin	barn	nya	fler	saring
april	kommer	sverige	nya	nbsp
klockan	ska	antal	nyheter	fraringn
läs	vill	många	svd	när
permalänk	amp	län	amp	ska
amp	skrev	får	stockholm	stockholm
mars	får	debatt	får	bilen

Table 4.8: The ten most common words in topic 6-10

Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
English words	?	?	HTML-outlook	Sport and culture
the	kommenterar	rekommendationer	color	annons
and	dag	tweets	background	stockholm
for	hemliga	annons	fontsize	amp
you	svd	foto	solid	sport
with	turkey	plus	sansserif	nyheter
that	amp	senaste	width	malmö
louis	näringsliv	vill	none	kultur
this	fund	mer	arial	webbtv
are	börsen	sverige	fontfamily	prenumerera
you	tar	feb	lineheight	rekommendationer

Table 4.9: The ten most common words in topic 10-15

Topic 11	Topic 12	Topic 13	Topic 14	Topic 15
Forums and months	?	?	News topics	Cities
jan	plus	vecka	april	norrköping
poster	säsong	kommenterar	publicerad	stockholm
visningar	avsnitt	amp	apr	testa
feb	aftonbladet	fler	sport	amp
startad	amp	näringsliv	katrineholm	quiz
trådar	nya	ska	amp	important
senaste	vecka	vill	uppdaterad	krönikor
amp	nöjesbladet	equity	lokal	söker
dec	mer	sicav	kultur	senaste
idag	fler	facebook	nöje	kommun

Table 4.10: The ten most common words in topic 16-20

Topic 16	Topic 17	Topic 18	Topic 19	Topic 20
Forum words	News words	Forum words	Small talk words	Small talk words
registrerad	publicerad	inlägg	ska	feb
medlem	bild	visa	lite	expressen
citera	februari	amp	bara	fler
anmäl	uppdaterad	idag	kommer	läs
gilla	amp	medlem	bra	visa
plats	listan	forum	får	mer
corsair	feb	senaste	vill	vill
senaste	kultur	meddelande	kanske	dela
asus	nöje	fler	också	annons
intel	läs	ämne	göra	expressense

In Table 4.7, Table 4.8, Table 4.9 and Table 4.10 the ten most important words for all topics are shown. A name for the topic is suggested except that for topics that are difficult to describe, a question mark is shown instead. Some of the topics seem to be directly related to the classes, which may explain the high accuracy of the models.

This shows that even though the LDA is unsupervised (the classes are not known to the algorithm) the algorithm was able to find topics that can quite easily separate the different types of websites.

4.5 Latent Dirichlet Allocation with word variables

To further improve the model, the set of 20 LDA topic variables and the set of word variables were both used as covariates in a multinomial logistic regression classifier. This combined model's accuracy rate is 99.70%, which is very good. Furthermore the number of variables in this model is nearly 100 less than the variables in the model with only the word variables.

Table 4.11: Contingency table of classification for multinomial logistic regression with merged data and LDA-variables

	Real label			
Classified as	Blog	News site	Forum	Total
Blog	7,613	24	8	7,645
News site	6	6,738	1	6,745
Forum	15	2	4,385	4,402
Total	7,634	6,764	4,394	18,792

Appendix D shows the variables that are used for this classification and it can be seen that the variables for blogs, news sites and forums are greatly reduced when the LDA-variables were in the model. The fact that the model chooses the LDA-variables for this regression is an indication of the good predictive qualities of these variables.

5 Analysis

In this section the results are analyzed and the models compared.

5.1 Challenges

There have been numerous challenges during this thesis work; most issues concerned the text processing. One of the problems has been with the coding of different web sites. Since Swedish contains letters that are not standard (å, ä, ö), the coding of these letters varies between websites. This problem was partly addressed by the tm package in R, but sometimes, when å, ä and ö are coded with other letters, some information is lost. For example there should be no difference between the words "varför" and "varfoumlr", but the coding makes them two different variables.

Another problem with having the texts in Swedish is that the tm package in R is mainly built for English, and for example stemming is not adequatly implemented for Swedish. Stemming is supposed to return the words to the root word, for example the words "walks" and "walking" should be stemmed to "walk". This technique is used to group words with similar basic meaning. The use of stemming is not unproblematic though because words like "marketing" and "market" are stemmed to the same word (market), but are not closely related. In Swedish, however, the stemming algorithm in R does not produce relevant results, and stemming is therefore not used in this thesis.

5.2 Comparison of the models

A comparison between the different combinations of models and data sets is shown in Table 5.1. It can be seen that the model with highest accuracy is the model combining the LDA-topics, the word variables and the HTML-variable with multinomial logistic regression and Lasso regularization. It can also be seen that the Naive

Bayes classifier performed significantly worse than the multinomial logistic regression. The disadvantage with Naive Bayes as a model is that it is built on the assumption that the variables are independent. Often this is a good approximation, but in this case this may be the reason why Naive Bayes is outperformed by multinomial logistic regression.

Any choice of model is directly dependent on the desired outcome of the model. If the primary aim is to get as high accuracy as possible, then the best choice is using the model combining LDA-variables, word variables, HTML-variable and multinomial logistic regression. Interestingly, using the LDA technique to create word-set topics and using those topics together with the original words allows multinomial regression to improve the fit to available data.

Another criterion could be to get a model that is easy to interpret and use; in this case the multinomial logistic regression with just the word variables would be a better choice. This model is easier to interpret and there are not as many advanced calculations in this model as there are in the LDA, which in turn means that the algorithm is quicker to both fit to new training data and to use to classify new data. When dealing with large data sets, the calculation time may be so large that the more accurate model is too slow for the model to deliver the results within available time.

From Table 5.1 it seems that the time dependence is of great importance. The model fitted with only the data from January 2013 does not make a good prediction of the data from April 2013. This means that the period of data extraction is important and that websites changes over time to some extent. The model with both datasets performs better than the model with only the data from January 2013. The reason for this could be due to the larger number of observation in the training sample, but it could also be due to the model being less time dependent. Either way, to get a good model the data should not be extracted at the same instant. Preferably the data should at least be extracted over a year to avoid seasonal variation in the models.

Table 5.1: All the models put together

Model	Data set	Variables	Accuracy
Naive Bayes	Merged data	Word variables and HTML-variable	84.33%
MLR with Lasso	Data 1	Word variables (with numbers) and HTML-variable	99.62%
MLR with Lasso	Data 1	Word variables and HTML-variable	99.61%
MLR with Lasso	Data 1 (training) and data 2 (test)	Word variables and HTML-variable	90.74%
MLR with Lasso	Merged data	Word variables and HTML-variable	99.64%
MLR with Lasso	Merged data	5 LDA-topics and HTML-variable	85.46%
MLR with Lasso	Merged data	10 LDA-topics and HTML-variable	95.67%
MLR with Lasso	Merged data	20 LDA-topics and HTML-variable	97.62%
MLR with Lasso	Merged data	20 LDA-topics, word variables and HTML-variable	99.70%

5.3 Data format

A problem in this study is the format of the data. Due to the web crawler, the HTML-code sometimes appears in the parts were there is only supposed to be site content (the text). There is also a problem with the time when the data is extracted from the websites, since all data were extracted over two short time periods, words like "April" and "March" appear more than they probably would have done if the data had been extracted evenly over a year.

Another problem with the quality of the data is the limited number of domains. The forums and news sites were each extracted from 17 domains, but have 10,369 (forums) and 14,299 (news sites) observations. This means that there is a higher risk of overfitting the model to these domains, which would not be seen from the accuracy rate of the test set. When the first multinomial logistic regression model was fitted, the quality of the dataset was even worse since it had fewer domains.

6 Conclusion

Using a Naive Bayes classifier to predict the data gives an accuracy rate of 84.33%. This is quite low for a text classifier and is therefore not considered satisfactory.

When using only the dataset from January 2013 in a multinomial logistic regression with Lasso, the classification accuracy is 99.61%. When the same parameters are used, but with the dataset from April 2013 as test set, the accuracy is only 90.74%. This means that there are quality problems with the first dataset either because of time dependence or because of overfitting due to there being too few domains in the dataset. When another multinomial logistic regression with both datasets merged (observations from both of the datasets in both training and test set) is fitted, it gives a better accuracy rate of 99.64%, which is quite good.

The thesis shows that LDA can successfully be used to summarize a large number of word-variables into a much smaller set of topics. A multinomial logistic regression model with 20 topic variables as covariates obtained an accuracy of 97.62 %, which is quite impressive considering that more than $6,000$ words were condensed into a mere 20 topics. When a multinomial logistic regression with the LDA-variables, the numlink HTML-variable and the word variables from both datasets was fitted, the classification accuracy was 99.70%. Most of the LDA-variables are chosen by lasso, which shows that these are important for the result. This is notable since the LDA-variables are based on the word-variables.

6.1 Further work

There are many ways to improve further or extend the classification of websites based on their functionality in the future.

One way would be to investigate how, as in this case a website

evolves over time, ie. how the content of the websites such as news sites changes topics, and forums discussions changes character. That the topic of different websites changes topic is explored by D. Blei et al. in 2012 [18], where they analyze change in topics over time. In this case it would be interesting to analyze if the content for the different functionality types changes in a similar way or if some of them are more consistent than others or if some changes are so typical for the functionality that the change in itself can be used for classification.

Another aspect that may be improved in further work is to make better use of the HTML information. In this case most of the variables were pure text variables and only one variable was created from the HTML-code. In previous experiments in this area, the text variables were weighted depending on where in the HTML-code they appear. This is an interesting area for further work, because it can improve the classification filters even further.

There is also a possibility of including more variables extracted from the HTML-code to get improved predictive power. For example the number of pictures, the complexity of the website etc could be taken into consideration. Another interesting study would be adding more text variables, instead of counting the words of the website. The models considered in this thesis are all built on the bag-of-word assumptions or unigrams. This means that the words are counted separately. Extending the unigram framework with covariates based on higher order n-grams would extract more information from the texts and could possibly improve the models.

The training and test data sets here are all taken from the same small set of domains. They are of course not overlapping, but since they are from the same domain there may be similarities that cannot be removed before the modelling. This makes it interesting to examine how the predictive accuracy would be affected if additional domains were used as testsets with the models developed in this thesis. As mentioned previously in this thesis, the sample would also benefit from being extracted throughout the year to remove seasonal

time dependence.

Another method that could make the classification filter better would be to only use word variables for words appearing in a dictionary. This would reduce the domain dependency by omitting names and words that are put together because of bad performance of the web crawler. This could be combined with a good stemming algorithm to improve the selection of word variables for the classification.

In this thesis only three categories of websites are considered, but there are more functionality classes, such as personal webpages and game pages. It might be interesting to try the models from this thesis on a larger data material with more categories to see if the models can handle that as well or if more advanced models are needed.

Bibliography

[1] Daniele Riboni. Feature selection for web page classification, 2002.

[2] Shaobo Zhong and Dongsheng Zou. Web page classification using an ensemble of support vector machine classifiers. *JNW*, 6(11):1625–1630, 2011.

[3] Christopher D. Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, USA, 1999.

[4] Ajay S. Patil and B.V. Pawar. Automated classification of web sites using naive bayesian algorithm, 2012.

[5] Aixin Sun, Ee-Peng Lim, and Wee-Keong Ng. Web classification using support vector machine. In *Proceedings of the 4th international workshop on Web information and data management*, WIDM '02, pages 96–99, New York, NY, USA, 2002. ACM.

[6] Majid Yazdani, Milad Eftekhar, and Hassan Abolhassani. Tree-based method for classifying websites using extended hidden markov models. In *Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, PAKDD '09, pages 780–787, Berlin, Heidelberg, 2009. Springer-Verlag.

[7] Oh-Woog Kwon and Jong-Hyeok Lee. Web page classification based on k-nearest neighbor approach. In *Proceedings of the fifth international workshop on on Information retrieval with Asian languages*, IRAL '00, pages 9–15, New York, NY, USA, 2000. ACM.

[8] Dou Shen, Zheng Chen, Qiang Yang, Hua-Jun Zeng, Benyu Zhang, Yuchang Lu, and Wei-Ying Ma. Web-page classification through summarization. In *SIGIR*, pages 242–249, 2004.

[9] Dou Shen, Jian-Tao Sun, Qiang Yang, and Zheng Chen. A comparison of implicit and explicit links for web page classification. In *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pages 643–650, New York, NY, USA, 2006. ACM.

[10] Christoph Lindemann and Lars Littig. Classifying web sites. In *International World Wide Web Conferences, WWW*, pages 1143–1144, 2007.

[11] E. Elgersma and M. de Rijke. Learning to recognize blogs: A preliminary exploration. In *EACL 2006 Workshop on New Text: Wikis and Blogs and Other Dynamic Text Sources*, 2006.

[12] Jerome H. Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2 2010.

[13] Trevor J. Hastie, Robert John Tibshirani, and Jerome H. Friedman. *The elements of statistical learning : data mining, inference, and prediction.* Springer series in statistics. New York, N.Y. Springer, 2009.

[14] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

[15] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[16] David M. Blei. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, 2012.

[17] Statsoft Inc. Naive bayes classifier. http://www.statsoft.com/textbook/naive-bayes-classifier/, 2013. Accessed: 2013-05-17.

[18] Chong Wang, David M. Blei, and David Heckerman. Continuous time dynamic topic models. *CoRR*, abs/1206.3298, 2012.

A Multinomial logistic regression with number variables

Table A.1: The chosen important variables in the multinomial logistic regression with the number variables included.

Blogs	News sites	Forums
123	numlinks	115
127	000	116
149	1995	210
154	1px	458
aktivitet	aftonbladet	8250
aktuella	annons	8594
andra	avhandling	899
ange	backgroundcolor	aktivitet
annan	**behöva**	**andra**
annonsera	billigast	användarnamn
användarnamn	byggde	**bland**
arbete	copyright	**blogg**
avstånd	drabbas	bluerayhd
backgroundcolor	erbjudande	bytes
band	fler	**både**
behövde	framträdande	calendar
beställ	**förklarar**	endast
bidra	göteborgsposten	eposta
blogga	hierta	fixar
blogg	**istället**	forum
bloggen	johansson	hitta

brott	knep	inlägg
både	kommentaren	**konkret**
började	kontakta	kosttillskott
copyright	kronor	mjukvara
dagen	kronstams	mjukvarubaserad
dagsläget	krönikörer	moderatorer
delicious	ladda	olästa
delvis	lediga	postade
ens	leif	r252
fick	leifby	**sajt**
fin	lhc	seagate
flyttade	linköping	skicka
for	live	startade
framför	mejl	startat
framöver	mejla	streacom
from	mera	svsgruppköp
färdigt	mest	terry
följa	mobilen	topic
förbi	**mystiska**	topp
förlorat	måndagen	totalt
försöka	nyyheter	tråd
galaxy	nyhetergpse	tråden
genom	orrenius	utskriftsvänlig
god	polisens	viv
hel	privata	ämnen
helst	pågick	ämnet
hittade	regissören	
hitta	reklamslogans	
härliga	rock	
hästarna	rysk	
igenom	sajt	

ikväll	servicefinder	
inloggning	sidans	
inne	siffrorna	
inse	skandalerna	
internet	skola	
istället	skriv	
kompetens	**skull**	
kontakta	snabbmatskedjorna	
köp	tipsa	
lagt	trio	
leifby	unika	
like	utskrift	
lite	**varann**	
lyckats	varför	
lyssna	vädertrafikhär	
lyssnade	växel	
låst	york	
massa		
möjlighet		
naturligtvis		
offentligt		
ofta		
olika		
ort		
ovan		
parti		
playstation		
promenad		
resten		
roligt		
senaste		

sidor		
siffrorna		
själva		
skapa		
skicka		
skolan		
skriva		
skrivit		
skylla		
slutet		
speciellt		
spelas		
startade		
storm		
stängt		
sök		
söka		
tempo		
tiden		
tipsa		
topp		
trackbacks		
trodde		
tydligen		
tänka		
underbart		
utställning		
vacker		
vanligt		
varför		
where		

visa		
you		
åkte		
åren		
ämnet		
ändå		
äntligen		
även		
öppet		

B Multinomial logistic regression with word variables and dataset from January 2013

Table B.1: Chosen variables/word by lasso for the dataset from 2013 with multinomial logistic regression.

Blogs	News sites	Forums
ahover	numlinks	aktivitet
aktivitet	ahover	**andra**
alltså	anledning	användarnamn
ange	annons	apex
annan	**attityd**	bilstereo
annat	avisited	**bland**
annonsera	beställ	**blogg**
attityd	**biggest**	**bloggare**
bara	**borås**	**bloggportalen**
behövde	boys	chromebook
beställ	byggde	community
besöka	cecilia	ddr
besökte	chefredaktör	**emma**
bidra	copyright	endast
biggest	drabbas	eposta
blogg	erbjudande	forum
blogga	etidning	forumets
bloggen	fler	**föreslås**
blogginläggen	framträdande	hitta
borås	**funderingar**	**html**
business	förväntningarna	inlägg
både	grupp	london

började	gunnar	lucida
copyright	göteborgsposten	låst
delicious	hierta	lösenord
delvis	**istället**	media
dessutom	johansson	memory
dressyr	**jävla**	mjukvara
dricka	kommentaren	moderator
duktig	krönikörer	moderatorer
ens	**kär**	nyhetsrubriker
fick	lediga	olästa
fin	legend	postade
fitness	leif	**sajt**
flyttade	leifby	skicka
for	lhc	smallfont
framför	linköping	startade
förbättra	live	startat
förlorat	läs	streacom
försvunna	läsning	support
försöka	**mamma**	sälj
förväntningarna	mejla	tborder
genom	mera	tele
gymnasiet	mest	tillbehör
helst	minheight	topic
html	niklas	topp
härliga	norrköpings	totalt
idrott	nyheter	tråden
ikväll	nyhetergpse	utskriftsvänlig
intressanta	pdf	ämnen
istället	placera	ämnet
jävla	privata	översikt
kombination	reklamslogans	

kommentaren	rock	
kuriren	rysk	
köp	sajt	
lagt	servicefinder	
large	sidans	
leifby	skandalerna	
like	skola	
lite	skriv	
lucida	**skull**	
lyckats	snabbmatskedjorna	
lyssna	ställer	
låst	sudoku	
massa	tas	
minheight	**taxi**	
möjlighet	tipsa	
möjligheter	**trött**	
naturligtvis	tweet	
nämligen	urval	
olika	utskrift	
ort	vind	
ovan	vädertrafikhär	
parti	växel	
pdf		
place		
promenad		
ringde		
roligt		
rubriker		
råkade		
rök		
röra		

semester		
servicefinder		
sidor		
själva		
skapa		
skicka		
skolan		
skriv		
skriva		
skrivit		
smallfont		
smycken		
snabbmatskedjorna		
speciellt		
spelas		
startade		
startat		
stängt		
sök		
söka		
tanke		
taxi		
tborder		
texten		
tidsfördriv		
tills		
tipsa		
topp		
trackbacks		
trodde		
trött		

tydligen		
underbart		
uppdaterade		
utskrift		
vanligt		
webben		
veckorna		
where		
viktminskning		
visa		
viss		
you		
åkte		
åren		
åtminstone		
ämnet		
ändå		
äntligen		
även		
önskar		
öppet		
översikt		

C Multinomial logistic regression with word variables and merged datasets

Table C.1: Chosen variables/word by lasso for the merged dataset with multinomial logistic regression.

Blogs	News sites	Forums
aftonbladets	ampamp	**numlinks**
aktiva	**annas**	aktivitet
aktivitet	annons	android
alltid	annonsera	annonser
alltså	anställd	användarnamn
amp	ansvarig	asus
and	backgroundfff	avancerad
andra	bero	begagnade
annons	besked	**bekräfta**
annonsera	bland	bevaka
annonserna	bortom	**bland**
ansvarar	casinokollen	bmw
användarnamn	cecilia	bosatt
bekräfta	center	community
bero	champagne	copyright
beställ	chefredaktör	core
blogg	displaynone	**day**
blogga	divfirstchild	debatterna
bloggen	dnse	delar
boken	drabbas	diverse
bort	epostadress	drivs
brott	erbjudande	**ekonomiskt**

bussen	etidningen	**emma**
började	finska	endast
center	fler	**erbjudande**
cool	floatnone	faq
copyright	flygande	forum
cykla	**fråga**	forumet
dagen	frågetecken	forumtrådar
dar	förhandlingar	fyll
debatter	förvandlar	fönster
drivs	galen	**förr**
suktig	**gissa**	försök
emma	göteborgsposten	**förändringar**
epostadress	hierta	galleri
fall	**hålla**	**gången**
fin	inled	**göteborgsposten**
finna	**istället**	**heter**
folket	**journalistik**	hitta
for	knep	htpc
framför	kommentar	hör
framöver	kommentaren	import
följa	krock	info
försöka	kronor	inloggad
förutom	kryssning	inlägg
föräldrar	krönikor	**intressanta**
galen	larsson	journalistik
genom	lediga	juridik
given	leftauto	**kommentarerna**
givetvis	lhc	**kommentera**
grund	life	kontakt
grått	liga	kontroll
hejsan	live	**krig**

hel	**lycka**	**känns**
here	magnetarmband	köra
hopp	mera	lagring
hoppas	mest	list
huvud	miljon	**lyckats**
huvudet	mobilsajt	**lyssna**
höra	niklas	låna
ihåg	nina	låst
ikväll	nyheter	**längst**
inloggad	nyhetergpse	**lätt**
inloggning	näthatet	lösenord
inlägg	**obs**	medlem
inne	persson	**människor**
instagram	plus	**naturligtvis**
internet	polisens	nbsp
intressanta	prenumerera	nintendo
intresserad	privata	nyhetsrubriker
istället	rad	**officiellt**
its	regissören	**opera**
just	resultat	playstation
jävla	**rights**	qualcomm
kanske	rock	radeon
kategorier	sedda	reg
kroppar	siffrorna	registrerad
k**ryssning**	skadad	relationer
kul	**skickar**	salu
kunna	smärta	shop
kvällen	sofia	skickar
leggings	starta	skruvar
like	startsida	smallfont
liten	storbritannien	**sociala**

looking	succé	startade
lägga	**support**	startat
lägger	svensson	streacom
länkar	sökte	**ständigt**
läsa	tas	stäng
mail	teater	stängt
massa	textstorlekminska	svar
mänskliga	tryck	tborder
nbsp	**tävla**	tele
now	ulf	**tjugo**
nyss	**ungdomar**	topp
näthatet	uppsala	totalt
oerhört	utgivare	tråd
oketgoriserad	utrikesminister	tråden
okej	utskrift	udda
orkar	website	ungdomar
otroligt	vind	utskriftsvänlig
part	väcker	**verk**
per	vädertrafikhär	vincent
playstation	värk	väggen
plötsligt	växel	**året**
promenad	york	ämne
prova	ändra	ämnen
qualcomm	**öppen**	ämnesverktyg
reg		ämnet
relationer		översikt
required		
resultat		
rights		
riktigt		
ryggen		

salu		
samla		
sedda		
sidor		
siffrorna		
skapa		
skapar		
skriv		
skriva		
skydda		
slags		
smallfont		
snygga		
sociala		
sovit		
spännande		
starta		
startade		
startsidan		
steget		
stod		
stängt		
svag		
säga		
säker		
säng		
sök		
tag		
tanke		
tas		
tborder		

textstorlek		
the		
tiden		
tillräckligt		
tills		
tipsa		
tjej		
topp		
totalt		
trackbacks		
tyskland		
underbara		
uppe		
ute		
vardag		
vare		
varenda		
webbläsare		
webbredaktör		
vilja		
vincent		
visa		
väl		
vänner		
väntade		
växel		
years		
åkte		
ändå		
även		
öppet		

överhuvudtaget		

D Multinomial logistic regression with LDA-topics and word variables

Table D.1: Chosen variables by the lasso in the multinomial logistic regression with both LDA-variables, HTML-variable and word variables.

Blogs	News sites	Forums
Topic 1	**Topic 2**	Topic 2
Topic 5	Topic 3	**Topic 3**
Topic 6	Topic 4	Topic 5
Topic 9	**Topic 6**	Topic 11
Topic 15	Topic 9	**Topic 15**
Topic 19	Topic 10	Topic 16
numlinks	Topic 12	Topic 18
aftonbladets	Topic 17	**Topic 20**
alltså	**Topic 19**	afrika
annons	Topic 20	aktivitet
annonsera	aktiviteter	annonser
ansvarar	ampamp	användarnamn
användarnamn	**annas**	asus
avslöjar	annons	avancerad
backgroundfff	bero	begagnade
blogg	besked	bevaka
blogga	birro	**bland**
bloggare	bland	bmw
center	casinokollen	bosatt
cool	cecilia	calendar
copyright	champagne	copyright

cry	**dar**	**dansk**
dagen	displaynone	debatterna
dansk	divfirstchild	delar
dar	dnse	diverse
endast	drabbas	drivs
english	**dras**	endast
enorma	**däck**	**erbjudande**
epostadress	epostadress	faq
folket	erbjudande	forum
fortsätta	etidning	forumet
framöver	etidningen	galleri
fönster	fler	**göteborgsposten**
försäljningen	floatnone	hampm
givetvis	frågetecken	import
grund	försvaret	info
hamnar	förvandlar	inloggad
kis	förväntningarna	inlägg
höger	galen	journalistik
hörde	göteborgsposten	juridik
ihåg	**hemsida**	kontroll
inloggad	hierta	lagring
inne	**hörde**	list
inrikes	idrott	**lån**
insett	inled	låst
internet	**istället**	lösenord
intressanta	knep	magasin
intresserade	kommentaren	mbit
istället	konst	**meddelanden**
its	kontaktinformation	medlem
july	krock	nbsp
jätte	kryssning	nyhetsrubriker

kategorier	**kräva**	**opera**
knep	krönikor	playstation
kommentaren	larsson	politik
kryssning	ledia	privatannonserna
kunna	leftauto	radeon
ledig	lhc	redaktionen
leggings	life	regler
looking	liga	saga
lägger	linköping	shit
länkar	**lycka**	shop
läsa	**lägga**	skickar
meddelanden	magnetarmband	**skrivit**
now	mejla	smallfont
nytta	mest	startade
nämligen	mobilsajt	startat
nätet	måndags	streacom
näthatet	negativt	ström
oerhört	nej	stängt
oftast	nina	svar
opera	**nummer**	svaret
otroligt	nyheter	tborder
per	nyhetergpse	tele
persson	**obs**	tillbehör
playstation	**oftast**	topic
prova	persson	topp
rensa	plus	totalt
resultat	prenumerera	tråd
rights	privata	tråden
riksdagen	regissören	udda
rättigheter	resultat	ungdomar
sedda	**rights**	utskriftsvänlig

sida	rock	**verk**
siffrorna	räddar	verktyg
själva	**saga**	vincent
skriv	sedda	ämne
skydda	sidans	ämnet
snygga	siffrorna	översikt
sociala	sju	
spelas	skadad	
startade	**skickar**	
startsidan	skolor	
ström	startsida	
stängt	startsidan	
synnerhet	storbritannien	
säker	succé	
säng	svensson	
sättet	sökte	
sök	tas	
sökte	teater	
tas	**tills**	
tipsa	**tävla**	
topp	**ungdomar**	
totalt	utgivare	
trackbacks	utrikesminister	
underbara	utskrift	
uppe	website	
usel	vind	
utrikesminister	väcker	
vare	vädertrafikhär	
varumärke	värk	
where	växel	
vincent	york	

visa	ändra	
välkommen	**öppen**	
vänligen		
växel		
åtminstone		
äger		
även		
öppen		
öppet		

Avdelning, Institution
Division, Department

Statistik, Institutionen för Datavetenskap
Statistics, Department of Computer and Information Science

Datum: Juni 2013
Date: June 2013

Språk
Language

☐ Svenska/Swedish
■ Engelska/English

☐ _____

Rapporttyp
Report category

☐ Licentiatavhandling
■ Examensarbete
☐ C-uppsats
☐ D-uppsats
☐ Övrig rapport

☐ _____

ISBN

ISRN **LIU-IDA/STAT-A--13/004—SE**

Serietitel och serienummer **ISSN**
Title of series, numbering

URL för elektronisk version
http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-93702

Titel **Functionality classification filter for websites**
Title

Författare **Lotta Järvstråt**
Author

Sammanfattning
Abstract

The objective of this thesis is to evaluate different models and methods for website classification. The websites are classified based on their functionality, in this case specifically whether they are forums, news sites or blogs. The analysis aims at solving a search engine problem, which means that it is interesting to know from which categories in a information search the results come.

The data consists of two datasets, extracted from the web in January and April 2013. Together these data sets consist of approximately 40.000 observations, with each observation being the extracted text from the website. Approximately 7.000 new word variables were subsequently created from this text, as were variables based on Latent Dirichlet Allocation. One variable (the number of links) was created using the HTML-code for the web site.

These data sets are used both in multinomial logistic regression with Lasso regularization, and to create a Naive Bayes classifier. The best classifier for the data material studied was achieved when using Lasso for all variables with multinomial logistic regression to reduce the number of variables. The accuracy of this model is 99.70 %.

When time dependency of the models is considered, using the first data to make the model and the second data for testing, the accuracy, however, is only 90.74 %. This indicates that the data is time dependent and that websites topics change over time.

Nyckelord **Website classification, Functionality, Latent Dirichlet Allocation, Multinomial logistic regression**
Keyword

LIU-IDA/STAT-A--13/004—SE