

Get Started: Blogs 101

-By Hemant Joshi

‡ This document is current as of Aug 23, 2006 and by no means is it complete.

† The words blogs and documents have been used interchangeably in this draft but they refer to the same thing here.

* Some knowledge of Machine Learning techniques is assumed here.

- ❖ Blogs are generally views expressed about certain entity by an individual or a group of individuals.
- ❖ Blog classification can be at various levels. For example,
 1. spam blogs (splogs) vs. normal blog classification
 2. Blog vs. non-blog classification
 3. Opinionated blogs vs. not-opinionated blogs classification
 4. Positively opinionated blogs vs. negatively opinionated blogs classification
- ❖ Typical classification algorithms are supervised classifier based, where training set is provided to the classifier algorithm to learn the model and determine the class label of a particular blog. Once acceptable accuracy is obtained in the training phase, test phase (remaining blogs from the dataset that were not used in training) begins. In the testing phase, classifier predicts the class labels. If the class labels for the test data are known, accuracy can be determined in this phase as well. If the test dataset is very large, random sampling is done to determine the accuracy. During training phase, cross-validation is often used (10-fold in general) to increase accuracy and avoid over fitting models.
- ❖ Among supervised classification algorithms, Naïve Bayes (multi-nominal) classifier is very fast and very simple algorithm. Naïve Bayes algorithm estimates posterior probability of a document belonging to a particular class. Even though the algorithm is fast, it is often criticized because of lack of balance between high precision as well as high recall. We need a classifier that will give correct classification for each document (High Precision) and also gives only the correctly classified documents (High Recall)
- ❖ Decision Tree algorithms like ID3 form decision trees to determine if the sequence of nodes (features) weighted at certain values will lead to one class label or another. Any supervised classification algorithm is as good as the training data used to represent the complete dataset. Decision trees can be very large if the number of features is in tens of thousands of nodes. Decision trees do achieve good performance have a strong statistical basis but the trees could be difficult to manage and load in memory. Other approaches have been suggested to improve the performance of the trees and even partial loading of trees is possible in some cases.

- ❖ Support Vector Machines (SVMs) are that family of classifiers that try to determine if the classes can be separated by a linear or non-linear space (known as boundary). For high dimensional data like text, the boundary is generally a hyperplane which separates the given N classes. Most blog classification problems require bi-class (N=2) classification. SVMs are highly precision oriented in the sense that they maximize the width of the boundary hyperplane separating the classes. Hyperplane boundary is determined and limited by the number of support vectors. Support vectors are those data points in the space that determine the equation representing the hyperplane. Data points in this case are individual blogs represented in higher dimensional space. In the testing phase, data points are projected onto model space obtained with boundary equation during training. By determining which side of the boundary data point is on, and how far the data point is from the boundary, SVM can predict the class label with certain confidence. Kernel is a component of SVM and is responsible for the model adaptations. Different kernels are supported in SVM that can separate the space into linear or non-linear sub-spaces. Typically linear kernels are known to perform well for text classification problems. One good feature of SVMs is that they work well independent of number of features and so are highly suitable for high dimensional data like text.

Next I will try to answer few questions regarding blogs and blog classification (IMHO).

- ❖ What is a blog?

It is very controversial issue in the research community at least as to what constitutes as a blog and what does not. Various experts have contradictory views on this subject. We will try to focus on what is known and widely accepted notion. Blogs are different from personal web pages and other type of web media in the sense that blogs represent personal expression of opinion or emotion rather than simple information itself. News articles are written with objective of fact reporting. The objective behind blogging is different. Many experts consider this as a social collaboration phenomenon. News articles that maintain objectivity while reporting a story do not constitute a blog but personalized (sometimes opinionated) and subjective view regarding a certain story, situation or entity may be considered as a blog.

- ❖ So does subjective content exclusively defines a blog?

Unfortunately the answer to this question is not straight forward. Blogs not only often have subjective content they also have a mechanism where people can leave comments and feedback thus making content subjective and collaborative. If the story is written by an author in subjective style, will that make it a blog? You tell me. Let us focus on what we definitely know is a blog. Personal diaries on the web (blogs are also called web logs) are often blogs. Personal diary is highly subjective thought process of the author written down in certain flow describing an event, emotion or opinion. The account of hitchhiker backpacking through Europe is

considered his personal account and thus can be considered as a blog. Subjective content indicates high likelihood of blog content. If the travel agency advertises for Europe travels with good customer experiences, you won't call it a blog. Perhaps, I should say it is mostly easy for humans to identify content and context easily. Which means humans are better at finding blog content than machine learning algorithms. Sometimes two individuals may not agree on blog or non-blog content and that is the gray area.

❖ Is opinionated content a blog?

Most likely. Personal opinions are blogs. But this is not the only criteria to determine the content as blog. Blogs can be informative and non-opinionated.

❖ What are spam blogs (splogs)?

Typical blogs are content by an individual or group of individuals expressing views or objective information about any entity. Spammers started exploiting this avenue to promote certain products. Also blogging websites that have higher pagerank (Google ranks web-pages with pagerank) are common targets by spammers to promote advertisements. This results in ranking such spam web-pages higher in search engine results. Normal blogs also contains advertisements but splogs contain only advertisements. Splogs also contains certain keywords and links to other web-pages to obtain higher pagerank automatically.

❖ Do adjectives and adverbs in the content indicate a blog?

From Natural Language Processing (NLP) point of view, adjectives (e.g, good, nice, excellent) and adverbs (e.g. suck, like, hate) and even certain words in general (e.g. love, experience) indicate subjective view. This is not always true with blogs. Consider example of product marketing webpage with a sentence –

Product 'A' is not like anything out there.

Technically above sentence is subjective and opinionated but it is not a blog. The views should be explicitly opinionated as well. Also comments on the views also indicate opinionated interaction. Consider the example,

Product 'A' is much better than product 'B'.

In this case the views expressed are explicit and opinionated. NLP techniques are sometimes less effective for blog related research. This is due to use of slang, deliberate typos (e.g. argh! , grrrrreeeat, grrrrr)internet chat lingo (e.g. IMHO, LOL)and other colloquial words (e.g. B'day)common on blogs.

❖ Are online products reviews or service reviews blogs?

Most likely. Websites like epinions.com, cnet.com or even amazon.com allow their users to post product reviews online. Product reviews are highly opinionated and generally have explicit expression of personal views.

- ❖ What can be good (?) features for text classification, especially blog classification?

Features are the variables of each data point that allow training and testing of classification algorithms. If you can classify oranges and apples based on shape as well as color of the fruit, then shape and color are the two features of classification. Good features are certainly the ones that contribute in decision making as to which class the particular document belongs to. For example, shape and color are good features for distinguishing apples from oranges but weight may not give correct indication of what fruit it is and thus not so useful feature. Different classification problems of text use different feature selection techniques. Broadly I consider that there are 3 types of features as far as text classification is concerned.

1. Bag of words as features

This is most commonly used feature of text based classification. It lists words that appear in each document. Rarely occurring words or common words (like a, an, the) can be filtered. Stemming of words is also done sometimes to refer to all forms of *swim*, *swimmer*, *swimming* etc. as one word *swim*. The value of feature can be binary (1 if word present else 0), frequency based (occurs 3 times implies value 3) or normalized frequency based (calculate local and global weights of all words and use as value the normalized product of local and global weights).

NOTE: - It is observed that for typical text classification problems, binary (0 or 1) values of features are easy to obtain and provide good accuracy. Using normalized or simple frequency based weighting technique do not improve the performance of the classification. They do have computational overhead though.

2. Important words or seed words as features

For each text classification problem, you might have a set of words which are good indicators of class labels. Such words often called as seed words can be used to classify text. For example, words such as auto, finance, sports in the text content clearly indicate the category of the content. Similarly in case of blogs, words like nice, good bad, hate, sucks, never indicate presence of subjective views and thus likelihood of opinionated content. Seed words are normally chosen manually depending on the type of text classification problem. Sometimes N-grams are also selected as seed words instead of simple unigrams (e.g. bad video, screen scratches) for specific problems.

3. Purely statistical features

In this category words of the document or seed words are not considered as features. But instead other statistical features of the corpus are studied and used. For example, no. of words per document or length of document in bytes or ratio of adjectives to the total no. of words in a document are examples of statistical features. Statistical features are independent of the text and thus independent of the language used. These features are hard to understand at times and are of generic nature. For blog classification techniques, no. of opinionated sentences, ration of

adjectives and adverbs to total no. of words per document, no. of subjective sentences per document may be good candidates. Detailed study has to be conducted to observe impact of any particular feature.

❖ Final note:

Once the features are established, we still need to limit the space of features to only those that are important in classification. Thus feature selection techniques and algorithms have been proposed. Thorough comparison of all feature selection techniques is necessary. Certain feature selection strategies may work well with a particular type of dataset or even particular type of classification algorithm. Classification is supposed to be only as good as the features used to train the data. Features selected and the training data used should be true representation of the actual data. Classifiers like SVM are independent of number of features but still depend heavily on type of features. Good feature selection is still a large problem at hand for many classification problems.

❖ Any other salient features of blogs?

Yes there are. Typically blogs are updated frequently than web-pages with information on products. Also blogs allow comments on comments to be expressed. Even though we consider text based blogging here, audio as well as video blogging is popular form of blogging too. Many bloggers link pictures and news articles on the web or any other media content they come across recently (<http://www.youtube.com>) Many corporations allow their developers and hired internet evangelists to write blogs promoting lifestyle, choice of products/services or similar. Many corporations have internal blogs that are not available to public. Hate blogs and love-blogs are sometimes part of propaganda or political campaign to reach voters surfing the web. Though English is the most popular language on blog, it is certainly not the only language. Community blogs managed by several individuals are also popular on the internet.

❖ To read

- <http://ebiquity.umbc.edu/blogger/>
- [Learning to classify text](#)
- [Introduction to text categorization](#)
- [libSVM](#)
- [Weka](#)
- [TREC](#)
- [Opinion Mining](#)

❖