Personal vs Non-Personal Blogs

Initial Classification Experiments

Erik Elgersma
edelgers@science.uva.nl

Maarten de Rijke
mdr@science.uva.nl

ISLA, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam

ABSTRACT

We address the task of separating personal from non-personal blogs, and report on a set of baseline experiments where we compare the performance on a small set of features across a set of five classifiers. We show that with a limited set of features a performance of up to 90% can be obtained.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software

General Terms

Algorithms, Measurement, Performance, Experimentation

Keywords

Blog classification, language modeling

1. INTRODUCTION

Reliable blog classification is an important task in the blogosphere as it allows researchers, ping feeds (used to broadcast blog updates), trend analysis tools and many others to separate, e.g., real blog content from blog-like content such as bulletin boards, newsgroups or trade market reports [1], to isolate spam blogs [3], to track developments in the blogosphere [5], or to identify specific blog genres as in [6]. We address the task of distinguishing between personal blogs and non-personal blogs, a type of classification that is useful, inter alia, for media analysis, reputation tracking, and for tasks such as the "blog distillation" task considered at TREC 2007 [4]. We take a blog to be personal if it is an online journal, diary-like, in which the blogger keeps a running account of his or her daily life and shares intimate thoughts and feelings with the reader.

We report on a set of initial baseline experiments aimed at determining how basic classification and features sets perform on this task. Starting with a basic feature set (consisting of frequent uni-grams), and we add more blog features and find that with limited feature engineering standard text classifiers are able to achieve up to 90% accuracy scores.

Below, we detail our experimental setup, describe the features used, and then report on our classification results, both for individual features and for sets of features.

SIGIR'08, July 20–24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.

2. EXPERIMENTAL SETUP

We hand labeled a set of 152 blogs (76 personal and 76 non-personal) randomly sampled from a blog collection that we created for earlier blog classification work; see [1] for details on the collection; we used the definition of *personal blog* as given in the introduction: an online diary-like journal, in which the blogger keeps an account of her daily life and shares intimate thoughts and feelings.

We compared 5 machine learners implemented in the Weka toolkit [7]: Naive Bayes, SVM, kNN, decision trees, decision tables, and used all of them with default settings. Evaluation was done using 10-fold cross-validation, using "percentage correct" and precision and recall as the metrics on which we report.

As features for our classification experiments, we considered the following.

LM Following the popularity of features derived from unigram language models for text classification, we compare the frequency of frequently used terms in general blog text with the frequency of those terms in personal blog text. The feature score is obtained by summing all the ratio scores of the terms in general blog text.

Pronouns We also consider a simple variation on this idea by considering the percentage of pronouns in a blog, assuming that personal blogs are likely to contain more personal pronouns than, e.g., news or political blogs. We use a fixed list of (personal, interrogative, demonstrative, indefinite, relative and reflexive) pronouns[1] and take the feature value to be the fraction of pronouns in the blog.

InLinks We assume that non-personal blogs are more likely to have a large number of incoming links than personal ones, and use the Technorati Cosmos API[2] to obtain this number.

OutLinks Acting on the observation that personal blogs often have link to sites of interest to the blogger, we also obtain the number of outgoing links of a blog using the Technorati Cosmos API.

Hosts For three hosts—BlogDrive, BlogSpot, and LiveJournal—we have a feature to indicate whether a blog is

[1] Obtained from *The Tongue Untied, A Guide to grammar, punctuation and style*, URL: `http://grammar.uoregon.edu/pronouns/pronouns.html`

[2] `http://technorati.com/developers/api/cosomos.html`.

Table 1: Blog classification results, for all features separately, and for two groups of features (*All* and *All* without the LM feature). The numbers indicated the percentage of correctly classified blogs.

	Percentage correct (%)								
	LM	Pronouns	InLinks	OutLinks	Hosts–BD	Hosts–BS	Hosts–LJ	All\{LM}	All
NaiveBayes	53.95	75.00	61.84	61.84	51.97	62.50	81.58	61.18	59.87
SVM	49.34	60.53	50.00	51.32	49.34	62.50	81.58	83.55	83.55
kNN	49.34	75.00	69.74	69.08	51.97	62.50	81.58	82.89	82.89
Decision tree	45.39	77.63	68.42	69.08	49.34	62.50	81.58	88.82	84.21
Decision table	47.37	77.63	69.08	69.08	49.34	62.50	81.58	90.13	90.13

hosted by it (as many blogs on these platforms are of a personal nature).

We ran the following classification experiments: all features separately, all features combined, and all features minus the language modeling feature (LM) combined.

3. RESULTS

Table 1 lists the results of our classification experiments, listing the percentage correct scores. We see that, by itself, and independent of the learner used, the LM feature performs rather poorly, below the 50% baseline for most classifiers, indicating that personal and non-personal blogs are not separable in terms of general language usage.

However, if we look at the language usage of specific parts of speech, i.e., of pronouns, we do see noticeable improvements over the 50% baseline, with scores up to 77%, showing that there is a marked difference in the usage of this particular type of "language."

Next we look at the InLinks and OutLinks feature. Here, we again see a marked improvement over the 50% baseline, but we do not see much of a difference in performance between the two features.

The three hosts features behave differently. Hosts-BD (indicating whether a blog is hosted by BlogDrive) is unhelpful, Hosts-BS (indicating whether a blog is hosted by BlogSpot) consistently improves over the baseline, Hosts-LJ (indicating whether a blog is hosted by LiveJournal) helps even more, leading to a performance of over 80% correctly classified.

Finally, we look at two sets of features: one in which all features are included, and one in which all features but the LM feature is used. We see that the best performance is achieved by leaving out the LM feature—apparently, it is simply too noisy to be of any use, especially for the NaiveBayes classifier. The performance of the Decision tree and Decision table classifier are comparable, and somewhat higher than the performance of the SVM and kNN classifiers, which in turn outperform the NaiveBayes classifier.

Table 2: Precision and recall figures for two groups of features

	All\{*LM*}		*All*	
	Prec. (yes/no)	Recall (yes/no)	Prec. (yes/no)	Recall (yes/no)
NaiveBayes	0.57/0.87	0.96/0.26	0.58/0.80	0.93/0.26
SVM	0.98/0.76	0.68/0.99	0.98/0.76	0.68/0.99
kNN	0.84/0.82	0.82/0.84	0.85/0.81	0.80/0.86
Decision tree	0.95/0.83	0.82/0.96	0.86/0.83	0.82/0.87
Decision table	0.94/0.87	0.86/0.95	0.94/0.87	0.86/0.95

In Table 2 we take a closer look at the precision and recall figures for the two groups of features (*All* and *All*\{LM}). We see that, on the whole, and except for NaiveBayes, the precision numbers are mostly higher than the recall numbers, and that the LM feature has a slightly negative impact on both recall and precision (for the Decision tree learner).

4. CONCLUSION

We addressed the task of separating personal from non-personal blogs and reported on the outcomes of a set of initial baseline experiments aimed at this task. Off-the-shelf learners, with a small set of well-chosen features are able to achieve up to 90% correctly classified scores.

We also found that a standard language modeling-based feature set is not helpful in distinguishing between personal and non-personal blogs, although narrowing the feature set down to pronouns only does lead to substantial improvements over the baseline.

In future work we want to extend the set of labeled blogs used in our experiments, and examine additional features. Moreover, we aim to integrate the classification scores into our retrieval engine for both blog post and blog search [2], on the assumption that for many professional users, non-personal blogs should receive a boost in ranking.

5. ACKNOWLEDGMENTS

This research was supported by the E.U. IST programme of the 6th FP for RTD under project MultiMATCH contract IST-033104, and by the Netherlands Organisation for Scientific Research (NWO) under project numbers 220-80-001, 017.001.190, 640.001.501, 640.002.501, STE-07-012.

6. REFERENCES

[1] E. Elgersma and M. de Rijke. Learning to recognize blogs: A preliminary exploration. In *EACL 2006 Workshop on New Text*, March 2006.

[2] B. Ernsting, W. Weerkamp, and M. de Rijke. The University of Amsterdam at the TREC 2007 Blog Track. In *TREC 2007 Notebook*, 2007.

[3] P. Kolari, A. Java, T. Finin, T. Oates, and A. Joshi. Detecting spam blogs: A machine learning approach. In *Proceedings AAAI 2006*, 2006.

[4] C. Macdonald, I. Ounis, and I. Soboroff. Overview of the TREC 2007 Blog Track. In *TREC 2007 Notebook*, pages 31–43, 2007.

[5] G. Mishne. Experiments with mood classification in blog posts. In *Style 2005*, 2005.

[6] H. Qu, A. L. Pietra, and S. Poon. Blog classification using NLP: Challenges and pitfalls, 2006. URL: `http://www.hongqu.com/publications/Blog_Classification_NLP.pdf`.

[7] I. H. Witten and E. Frank. *Data Mining*. Morgan Kaufmann, 2nd edition, 2005.