Exploring Linguistic Features for Web Spam Detection: A Preliminary Study

Jakub Piskorski
Joint Research Centre
of the European Commission
Via Fermi 1
21020 Ispra, VA, Italy

Marcin Sydow
Polish-Japanese Institute
of Information Technology
Koszykowa 86
02-008 Warsaw, Poland

Dawid Weiss
Poznań University
of Technology
Piotrowo 2
60-965 Poznań, Poland

ABSTRACT

We study the usability of linguistic features in the Web spam classification task. The features were computed on two Web spam corpora: *Webspam-Uk2006* and *Webspam-Uk2007*, we make them publicly available for other researchers. Preliminary analysis seems to indicate that certain linguistic features may be useful for the spam-detection task when combined with features studied elsewhere.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Linguistic processing*; I.2.6 [**Artificial Intelligence**]: Learning

General Terms

Web spam

Keywords

Web spam detection, content features, linguistic features

1. INTRODUCTION

In this paper, we extend the work reported in Sydow et al. [12] by introducing more linguistic-based features and studying their potential usability for Web spam classification. Our effort is complementary to the work on content-based features reported by others (see Section 1.1). The main contributions are: (1) Computing over 200 new linguistic-based attributes; in order to get a better, less biased insight, we tested various NLP tools and two Web spam corpora together with 3 different document length-restriction modes. (2) Preparing and studying over 1200 distributions of all the attributes as potential Web spam discriminators. (3) Experimentally identifying the most promising attributes with use of 2 objective metrics. (4) Making the computed attributes with the corresponding histograms available for the research community at: `http://www.pjwstk.edu.pl/~msyd/lingSpamFeatures.html`.

AIRWeb '08, April 22, 2008 Beijing, China.

All the features discussed in this paper are computationally amenable and can be calculated at the document level which can be beneficial for future on-line computation thereof.

1.1 Related Work

The usability of various content-based features for the successful Web spam classification has been reported before. Fetterly et al. [7] proved that simple frequency-based measures are useful for the task. Drost et al. [4] extended the list by adding features based on checksums and word weighting techniques. This direction was further explored by Ntoulas et al. in [11]. Mishne et al. [10] analyzed the contents of a Web document to compare its language model with that of the citing blog in order to detect *blog spam*. Fetterly et al. [8] reported on techniques for identifying spam pages, whose content is automatically generated by gluing together phrases copied from non-spam pages. Urvoy et al. [13] introduced features based on HTML document structure to detect automatically generated, pattern-based spam pages. Benczur et al. [2] studied commercial attractiveness of pages via utilization of *Microsoft OCI*, *Yahoo! Mindset*, and keywords extracted from *Google AdWords* and *Google AdSense*.

Building on many previous results, some authors reported automatic spam classification utilizing both content-based and link-based features and deploying additional more sophisticated techniques on top of these features including bagging, exploration of the link structure for label-smoothing, 2-level stacked graphical learning, and graph regularizations [3, 1].

2. LINGUISTIC FEATURES

Zhou et al. [15] proved that certain language features have discriminatory potential for human deception detection in text-based communication. Motivated by this work, we selected and adapted a subset of these features for analyzing their usability for Web spam classification. We considered only features, whose computation does not involve much linguistic sophistication since the open and unrestricted nature of texts on the Web indicates that utilization of any more error-prone higher-level linguistic tools would introduce more noise.

Two NLP tools were used to compute linguistic features: *Corleone* [9], which comes with a morphological analyzer based on the extended MULTEXT resources [5][1], and *General Inquirer*[2]—a tool which maps an input text with counts

[1]It has an average coverage of 95% on unseen data
[2]`http://www.wjh.harvard.edu/~inquirer`

on dictionary-supplied categories (performing also word sense disambiguation). The current version combines the *Harvard IV-4* and *Laswell* dictionary content-analysis categories, totalling 182 categories in all.

We limited computing the features for HTML bodies of each page (converted to text) and made a simplistic assumption that all the processed texts are in English, which seems to be true for most of the documents in the `.uk` domain.

2.1 Corleone-based features

The features computed with *Corleone* are mainly based on statistics of part-of-speech (POS) information. However, no POS disambiguation has been performed for the reasons mentioned earlier (open-domain character of the Web)—when we refer to POS tags here, in case of ambiguous words, the tags represent all readings, e.g., the word *fight* is assigned the `NV` tag since it could be either a noun (`N`) or a verb (`V`).

Type: Web pages may include free text, numerical data or a combination of both. We introduced two attributes to estimate the 'type' (character) of the page:

$$\textit{Lexical validity} = \frac{\#\text{ of valid word forms}}{\#\text{ of all tokens}}$$

$$\textit{Text-like fraction} = \frac{\#\text{ of potential word forms}}{\#\text{ of all tokens}}$$

The term 'potential word forms' refers to the tokens which undergo morphological analysis—tokens representing numbers, URLs, punctuation signs and non-letter symbols are not counted as potential word forms.

Quantity: In the context of text quantity we computed the ratio of nouns (*Noun Fraction*), verbs (*Verb Fraction*), pronouns (*Pronoun Fraction*) and tokens starting with capital letters (*Capitalized Tokens*) to the total number of tokens.

Diversity: We have explored three types of text diversity, namely lexical diversity, content diversity and syntactical diversity, which are defined as follows.

$$\textit{Lexical diversity} = \frac{\#\text{ of different tokens}}{\#\text{ of all tokens}}$$

$$\textit{Content diversity} = \frac{\#\text{ of different nouns \& verbs}}{\#\text{ of all nouns \& verbs}}$$

$$\textit{Syntactical diversity} = \frac{\#\text{ of different POS n-grams}}{\#\text{ of all POS n-grams}}$$

The words with an initial capital, which were tagged by the morphological component as 'unknown' were considered in the context of computing *Content diversity* as nouns. Syntactical diversity has been calculated for 2, 3 and 4-grams.

Further, we computed *Syntactical entropy*, i.e., the entropy of the distribution of POS-based n-grams (2, 3 and 4 grams). Let $G = g_1, \ldots, g_k$ be the set of all POS n-grams in a page and let $\{p_g\}$ be the distribution of POS n-grams in G. The syntactical entropy is calculated as:

$$\textit{Syntactical Entropy} = -\sum_{g \in G} p_g \cdot \log p_g$$

Expressivity: As an indication of language expressivity, we have selected *Emotiveness*, which is the ratio of modifiers to content words, i.e., it is formally defined as follows.

$$\textit{Emotiveness} = \frac{\#\text{ of adjectives \& adverbs}}{\#\text{ of all nouns \& verbs}}$$

Non-immediacy: Linguistic non-immediacy can be seen as the degree of verbal indirectness with which communicators refer to themselves. We defined two scores for measuring the degree of non-immediacy, which are defined as follows.

$$\textit{Passive Voice} = \frac{\#\text{ of passive constructions}}{\#\text{ of all verbs}}$$

$$\textit{Self Referencing} = \frac{\#\text{ of 1st person pronouns}}{\#\text{ of all pronouns}}$$

Uncertainty: For measuring the uncertainty we computed the ratio of modal verbs to the total number of verbs in a page (*Modal Verbs*).

Affect: For computing the affect of pages we have utilized SentiWordNet [6] (integrated in *Corleone*), in which each synset s of WordNet[3] is associated with two numerical scores, namely, the *Pos(s)* and *Neg(s)*, which describe how 'positive' and 'negative' the terms contained in synset s are. In particular, we computed *PosSent* and *NegSent* score of text $T = t_1 \ldots t_n$, where t_i is the i-th token in the text, as follows.

$$\textit{PosSent} = \frac{\sum_{t \in T} \max(\textit{Pos}(t))}{\sum_{t \in T} \max(\textit{Pos}(t)) + \max(\textit{Neg}(t))}$$

$$\textit{NegSent} = \frac{\sum_{t \in T} \max(\textit{Neg}(t))}{\sum_{t \in T} \max(\textit{Pos}(t)) + \max(\textit{Neg}(t))}$$

A token (term) t might potentially have different senses. Therefore, we compute for each token t the maximum of *Pos* and *Neg* scores for the corresponding synsets t belongs to. In this way, terms, which might be used in both positive and negative sense of equal strength, are somehow 'neutralized'.

Next, we computed an unweighted affect score, which is the ratio of 'positive' tonality expressions to all opinion expressions (*Tonality*) based on ca. 1200 opinion expressions.

- 'Osgood' semantic dimensions
- pleasure, pain, virtue and vice
- overstatement/understatement
- language of a particular 'institution'
- roles, collectivities, rituals, and interpersonal relations
- references to people/animals
- processes of communicating
- valuing of status, honor, recognition and prestige
- references to locations
- references to objects
- cognitive orientation
- pronoun types
- negation and interjections
- verb types
- adjective types
- skill categories
- motivation
- adjective types
- power
- rectitude
- affection
- wealth
- well-being
- enlightenment

Figure 1: Overview of *GI* categories.

2.2 Features obtained with General Inquirer

General Inquirer (*GI*) 182 categories were developed for social-science content analysis applications. The values assigned by *GI* for these categories are based on occurrence statistics. Some of them overlap with the features defined in section 2.1. We treated each *GI* category as a separate feature. A snapshot of category types covered by *GI* is given in Table 1, more details are available at the Web page.[4]

3. EXPERIMENTS

3.1 Data sets

We used two Web spam data sets: *WebSpam-Uk2006* and *WebSpam-Uk2007* [14]. These data sets concern two general crawls of the `.uk` Web domain and provide each page's

[3] `http://wordnet.princeton.edu`

[4] `http://www.wjh.harvard.edu/~inquirer/homecat.htm`

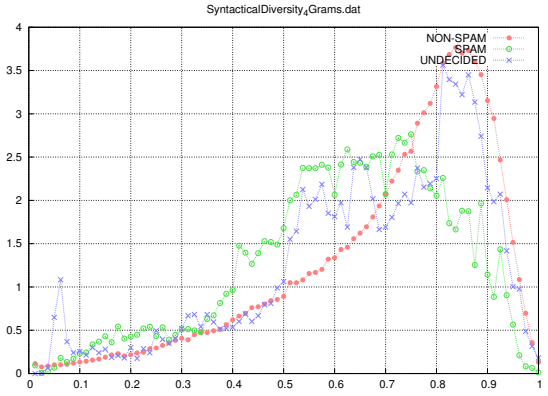

Figure 2: The histograms for *Syntactical Diversity* (4-grams), mode-2, *WebSpam-Uk2007*.

content, links and human-assessed categories of each host (spam, non-spam, borderline, undecided).

In our experiments, we restricted ourselves to a breadth-first 400-page-per-host sample of the overall collection. The original GZIP-compressed WARC files were first converted into binary, block-compressed sequence files, suitable for distributed parallel processing using the map-reduce programming paradigm. The Hadoop project[5] running on a cluster of 10 quad-core machines was used for all computations. The data set statistics are given in Table 1.

3.2 Histograms

For the two data sets we calculated linguistic attributes using *Corleone* and *General Inquirer*. We considered 3 different modes: (0) all the non-empty documents containing less than 50k tokens, (1) only the documents containing between 150 and 20k tokens and (2) only the documents containing between 400 and 5k tokens. The modes (1) and (2) were introduced to examine the influence of very short documents (which potentially bring much noise) and very long documents (which slow down the computations without adding much information, due to the power-law-like distribution of the document lengths).

For each of the 6 setting combinations described above we computed 208 linguistic attributes (23 for *Corleone* and 185 for *General Inquirer*) for each page satisfying the length constraints of the appropriate mode. Then, for each attribute's value we created document-level distribution histograms for the three possible human-given labels: spam, non-spam, borderline (documents receive the label of their host, since no document-level labels are available for either *WebSpam-Uk2006* or *WebSpam-Uk2007* corpus). This resulted in over 1200 histograms. All histograms represent attributes' value ranges (bins) on the horizontal axis and the percentage of pages that fell in each bin on the vertical axis. A preliminary study of the histograms revealed that:

- the histograms generated with the length restrictions are noticeably less noisy than the corresponding histograms computed without such restrictions (mode 0)
- mode-2 histograms seemed to be a bit less noisy than mode-1 histograms. Therefore, we used only the mode-2 attributes for computing the statistics in the next section.
- the borderline histograms are in general closer to spam histograms, but this has to be studied more rigorously.

[5] `http://hadoop.apache.org`

Table 1: Simple statistics of input data sets.

	2006	2007
pages	3 396 900	12 533 652
pages without content	65 948	1 616 853
pages with HTTP/404	281 875	230 120
WARC (compressed file, GB)	14.10	45.50
HTML SQF (compressed file, GB)	11.70	37.80
TXT SQF (compressed file, GB)	2.87	8.24

3.3 Objective selection of the best attributes

To extend the subjective observations and preliminarily identify the most promising attributes with some objective methodology, we introduced two difference measures: *absDist* and *sqDist*, defined below. For each attribute we measured the difference between spam and non-spam class distributions using the measures.

Let, for some attribute histogram h, $\{s^h\}_i$ and $\{n^h\}_i$ denote the sequences of heights of the bars for spam and non-spam classes, respectively, for all the considered bins $i \in I$. We define the distance metric *absDist* as follows:

$$absDist(h) \quad = \quad \sum_{i \in I} |s_i^h - n_i^h| / 200$$

which can be interpreted as the fraction of the total area under the histogram curves corresponding to the symmetric difference between them (the area under each histogram is equal to 100 units). Another distance measure is *sqDist*, defined for a histogram h as:

$$sqDist(h) = \sum_{i \in I} (s_i^h / max_h - n_i^h / max_h)^2 / |I|$$

where max_h, a kind of normalization factor, is defined as the maximum value among both $\{s^h\}_i$ and $\{n^h\}_i$.

The first metric seems to be more intuitive, since it has more natural geometrical interpretation, however using two different metrics may result in less bias. For both metrics the higher values indicate better discriminative power.

Next, for each out of over 1200 histograms we computed both measures and, for each of the 6 settings, we sorted the values decreasingly to identify topmost attributes. Interestingly, the choice of the length-restriction mode is almost insignificant to the ranking of the top-10 attributes in any of the 6 settings for both distance measures. For example, the list of the top 9 *Corleone* attributes (according to any distance measure) is the same for both data sets despite the fact that the 2 metrics are quite different. The results are presented in Table 2. In case of *abstDist* some histograms differ by almost 25% of the AUC, which seems to indicate their potential discriminatory power. The histograms for *Syntactical Diversity* with 4-grams for the mode-2 size-filtered *WebSpam-Uk2007* corpus is shown in Figure 2.

We did analogous experiments also for the *GI*-generated attributes. For the *absDist* metric, the list of the top-7 attributes was identical on both the data sets (though the ordering was a bit different) (see Table 3). Notice that for some attributes the histograms for spam and non-spam differ by almost 30% of the AUC, which is even more promising than in the case of the *Corleone*-generated attributes.

The *sqDist* metric identified two identical top-9 lists of attributes generated by *General Inquirer* for both corpora

Table 2: The most discriminating *Corleone* attributes wrt *absDist* and *sqDist* metric.

Corleone(absDist)	2007	2006	Corleone (sqDist)	2007	2006
Passive Voice	0.263	0.273	Syn. Diversity (4g)	0.053	0.054
Syn. Diversity (4g)	0.255	0.245	Syn. Diversity (2g)	0.050	0.067
Content Diversity	0.234	0.331	Syn. Diversity (2g)	0.037	0.036
Syn. Diversity (3g)	0.230	0.253	Content Diversity	0.032	0.065
Pronoun Fraction	0.224	0.261	Syn. Entropy (2g)	0.029	0.026
Syn. Diversity (2g)	0.221	0.232	Lexical Diversity	0.026	0.043
Lexical Diversity	0.213	0.262	Lexical Validity	0.024	0.033
Syn. Entropy (2g)	0.208	0.179	Pronoun Fraction	0.024	0.031
Text-Like Fraction	0.188	0.184	Text-Like Fraction	0.023	0.017

(Table 3). There is a significant overlap between the 2 lists of attributes identified by the two, quite different metrics.

Some *GI*-generated attributes were identified as among the top-10 by one or another measure or setting. Among them were: *EnlOth*, *EnlTot* (enlightenment words) *WltTot*, *WltOth* (words in wealth domain, e.g., pursuit of wealth), *ECON*, *Econ@dat* (words of commercial or business orientation), *Objects* (words referring to objects, e.g., food, vehicles, buildings), and *Leftovers* (encompasses several 'Lasswell dictionary' attributes not associated with any other 'large' categories in *GI*, and includes words of accomplishment, transaction, desired or undesired ends or goals, words referring to means and utility, words denoting actors, nations and emotions). The fact that the top *GI*-generated attribute lists are less stable than the *Corleone*-generated ones, can be explained by the fact that the size of the full set of *GI* attributes is over 7 times bigger than the *Corleone* set).

Table 3: The most discriminating *General Inquirer* attributes according to *absDist* and *sqDist* metric.

GI(absDst)	2007	2006	GI(sqDist)	2007	2006
WltTot	0.287	0.346	leftovers	0.0150	0.0128
WltOth	0.285	0.341	EnlOth	0.0085	0.0072
Academ	0.270	0.263	EnlTot	0.0082	0.0118
Object	0.255	0.282	Object	0.0073	0.0086
EnlTot	0.249	0.247	text-length	0.0056	0.0048
Econ@	0.228	0.356	ECON	0.0038	0.0034
SV	0.206	0.260	Econ@	0.0038	0.0031
			WltTot	0.0038	0.0027
			WltOth	0.0037	0.0024

3.4 Discussion

In general, we observed that the best attributes that we have computed (accordingly to the applied metrics) are quite promising for discriminating between the spam and non-spam classes, due to quite remarkable distribution differences. Also, the top-lists of the attributes are quite stable wrt the choice of the Web Corpus, which is an important positive property of the studied attributes.

The tables presented in the previous section seem to indicate that some *GI*-generated attributes have more potential discriminative power than *Corleone*-generated ones. Unsurprisingly, the top-scoring *GI* attributes refer to vocabulary centering around purchasing goods, transactions, business, industry, non-human objects and enlightenment.

Another interesting finding is that syntactical diversity showed better discriminative power than lexical diversity. This means that spam pages consisting of loosely assembled keywords can be determined using shallow syntactical analysis. On the other hand, no attribute showed clear separation between spam and non-spam classes. This is most likely because of the fact that spammers often reuse existing Web content and either repeat it literally or interleave it with their own content. Finally, in contrast to the work presented in [15], expressivity, uncertainty, and affect, do not seem to have any discriminatory power for differentiating spam from non-spam pages. Possibly, more sophisticated attributes for computing the aforementioned language features should be studied in order to get a better insight.

4. CONCLUSIONS

We reported on the computation of over 200 linguistic-based attributes on two publicly available reference Web spam corpora and discussed the general properties of over 1200 analyzed histograms.[6] To our best knowledge, such attributes have not been previously studied in the context of Web spam detection. In particular, two distribution difference metrics were used for identifying the most promising attributes. The list of the latter ones seems to be stable across a couple of different settings, however their real usefulness is to be studied in the future. All the computed attributes and histograms are available for research purposes.

5. REFERENCES

[1] J. Abernethy, O. Chapelle, and C. Castillo. Witch: A new approach to web spam detection, 2007. submitted.

[2] A. Benczúr, I. Bíró, K. Csalogány, and T. Sarlós. Web spam detection via commercial intent analysis. In *Proceedings of AIRWeb 2007*, pages 89–92, New York, NY, USA, 2007. ACM.

[3] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri. Know your neighbors: web spam detection using the web topology. In *SIGIR '07: Proceedings of the 30th ACM SIGIR conference, Amsterdam, The Netherlands*, pages 423–430. ACM, 2007.

[4] I. Drost and T. Scheffer. Thwarting the nigritude ultramarine: learning to identify link spam. In *Proceedings of ECML 2005*, volume 3720 of *LNAI*, pages 233–243, Porto, Portugal, 2005.

[5] T. Erjavec. MULTEXT – East Morphosyntactic Specifications, 2004. URL: `http://nl.ijs.si/ME/V3/msd/html`.

[6] A. Esuli and F. Sebastiani. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC 2006*, pages 417–422, Genova, IT, 2006.

[7] D. Fetterly, M. Manasse, and M. Najork. Spam, damn spam, and statistics: using statistical analysis to locate spam web pages. In *Proceedings of WebDB '04*, New York, USA, 2004.

[8] D. Fetterly, M. Manasse, and M. Najork. Detecting phrase-level duplication on the world wide web. In *Proceedings of SIGIR '05*, pages 170–177, New York, NY, USA, 2005. ACM.

[9] Jakub Piskorski. Corleone - Core Linguistic Entity Extraction. Technical Report. JRC of the European Commission, 2008.

[10] G. Mishne, D. Carmel, and R. Lempel. Blocking blog spam with language model disagreement. In *Proceedings of AIRWeb 2005*, May 2005.

[11] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly. Detecting spam web pages through content analysis. In *Proceedings of WWW 2006, Edinburgh, Scotland*, pages 83–92, 2006.

[12] M. Sydow, J. Piskorski, D. Weiss, and C. Castillo. Application of machine learning in combating web spam, 2007. submitted for publication in IOS Press.

[13] T. Urvoy, T. Lavergne, and P. Filoche. Tracking web spam with hidden style similarity. In *AIRWeb 2006*, pages 25–31, 2006.

[14] Webspam corpora. URL: `http://yr-bcn.es/webspam/datasets`, accessed February 21, 2008.

[15] A. Zhou, J. Burgoon, J. Nunamaker, and D. Twitchell. Automating Linguistics-Based Cues for Detecting Deception of Text-based Asynchronous Computer-Mediated Communication. *Group Decision and Negotiations*, 12:81–106, 2004.

[6] We would like to thank A. Wawer for directing us to *General Inquirer*, and all the contributors of the Hadoop project for their hard work. We used the HPC cluster at Poznan University of Technology for computations. The work presented in this paper was supported by the PJIIT research grant ST/SI/06/2007.