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

In this paper, we study the usability of linguistic features in the context of statistical-based machine-learning approach to the Web Spam detection problem. We give a brief description of these features and report on their computation for the two publicly available web-spam corpora, namely *Webspam-Uk2006* and *Webspam-Uk2007*. A preliminary analysis of multiple histograms revealed that some of the examined features exhibit potential usability for discriminating between spam and non-spam. In particular, we believe that some of them may be useful for the spam-detection task when combined with other known features studied elsewhere. We made the computed features together with the corresponding histograms publicly available for other researchers.

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

Due to the dominating role of search engines in accessing the on-line information and the existence of on-line advertising which makes it possible to directly correlate incomes with Web visibility, omni-present Web spam, as being highly economically motivated, seriously threatens the mission of search engines and the quality of information in the Web. Thus, improving the tools for automatic Web spam detection remains the top-importance issue and has recently gained increasing attention from the machine-learning community.

In this paper, we report preliminary analysis of using linguistic attributes as potential discriminators in the context of statistical-based, machine-learning approach to Web Spam detection. We believe that these attributes may improve the existing spam-detection systems by combining them with the other features successfully used before.

1.1 Related Work

The usability of various content-based features for the successful Web spam classification has been reported before. Fetterly et al. [9] proves that host name length, word number variation within a site, their frequency and extent of page content revisions are useful for the task. Drost et al. [6] extended the list by adding features based on checksums and word weighting techniques. Mishne et al. [12] 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. [10] report on techniques for identifying spam pages, whose content is automatically generated by glueing together phrases copied from non-spam pages. Ntoulas et al. [13] explore several new and useful content-based features, including the number of words in a page, page title length, average word length, fraction of anchor text, page compressibility, fraction of the page drawn from 'popular' words, fraction of 'most popular' words that appear in the page, and token-based n-gram likelihoods. Urvoy et al. [15] introduce features based on HTML document structure to detect automatically generated, pattern-based spam pages.

Building on many previous results, Castillo et al. [5] presented a multi-level process of automatic spam classification which used over 200 combined content-based, link-based and even query-log-based features. Additional techniques were employed on the top of these features: bagging, exploration of the link structure for label-smoothing and 2-level stacked graphical learning. Recently, [2] reported further work with link structure using graph regularization and support vector machines, resulting in outstanding performance measures (over 90% AUC measure).

Benczur et al. [3] studied features in the context of their commercial attractiveness. In particular, commercial intention of pages has been computed via utilization of *Microsoft OCI* (On-line Commercial Intention), *Yahoo! Mindset* classifier, and keyword frequency information extracted from *Google AdWords* and *Google AdSense*.

1.2 Contribution

In this paper, we extend the work reported in Sydow et al. [14] 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. According to our best knowledge, utility of content-based linguistic features to detect Web spam has not been studied before. The main contributions are described in the list below.

- 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.
- Preparing and studying over 1200 distributions of all the attributes as potential discriminators in Web spam classification.
- Experimentally identifying the most promising attributes with use of 2 objective metrics.
- The features computed with all the described modes and tools, accompanied with the corresponding histograms, are made available (under mild conditions) at `http://www.pjwstk.edu.pl/~msyd/lingSpamFeatures.html`

Our aim was to study the distributions of pure and generic linguistic features, which can be computed without any prior analysis of the whole web-spam training corpus. All the linguistic features discussed in this paper are computationally amenable and can be calculated at the document level, i.e., we did not utilize corpus statistics for calculating feature values in any way. This can be beneficial for future on-line computation of the features. Due to their known poor performance on open domains such as WWW we abandoned higher-level linguistic processing tools (full parsing).

We hope that the attributes presented here, in combination with other attributes, may be useful for improving spam classifiers.

2. LINGUISTIC FEATURES

There are many aspects that can be measured and extracted from the text apart from word statistics. Zhou et al. [16] proved that language features, such as expressivity, complexity, affect, informality, uncertainty, non-immediacy, diversity and emotional consistency, have discriminatory potential for human deception detection in text-based communication. Intuitively, their utilization for differentiating Web spam from legitimate content might be beneficial and, to our best knowledge, they have not been exploited in this context.

For our experiments, we selected and adapted a subset of the feature definitions described in [16] and added some new ones. Noteworthy, we consider here only the computationally amenable features, whose computation does not involve much linguistic sophistication. The open, unrestricted nature of the texts on the web indicates that utilization of any higher-level linguistic analysis tools, which are known to be more error prone, makes little sense. Consequently, the features explored in this paper can be seen as merely approximations of what could possibly be computed more precisely using some more sophisticated linguistic tools.

Two NLP tools were used to compute linguistic features: *Corleone* [11], which comes with a morphological analyzer based on the extended MULTEXT resources [7] (with an average coverage of 95% on unseen data), and *General Inquirer*[1] — a tool which maps an input text with counts on dictionary-supplied categories (performing word sense disambiguation not just dictionary look-up). 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). Since the NLP tools described above are only capable of processing English texts, we 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. It is important to note at this stage that no part-of-speech disambiguation has been performed for the reasons mentioned earlier (open, unrestricted character of the Web), i.e., when we refer to POS tags here, in case of ambiguous words, the tags represent all readings, e.g., the word *fight* is assigned `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 ‘number of potential word forms’ refers to the number of tokens which undergo morphological analysis—tokens representing numbers, URLs, punctuation signs and non-letter symbols are not counted as potential word forms.

Quantity: Text quantity has been already utilized as a content-based feature in [13]. We computed some statistics based on POS information, in particular, the ratio of nouns (*Noun Fraction*), verbs (*Verb Fraction*), and pronouns (*Pronoun Fraction*) to the total number of words in a page. Nouns and verbs are major content words. Further, we also computed the fraction of words in a page starting with a capital letter (*Capitalized Tokens*).

Diversity: Text diversity can be measured in several ways, e.g., as compression ratio [13]. We have explored three types of text diversity, namely lexical diversity, content diversity and syntactical diversity, which are defined as follows.

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

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

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

Note that all words with an initial capital letter, which have been tagged by the morphological component as ‘unknown’ (unrecognized) were considered in the context of computing *Content diversity* as nouns (content words). Syntactical diversity score 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:

[1] `http://www.wjh.harvard.edu/~inquirer`

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

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: The uncertainty can be measured in different ways, e.g., usage of modifiers makes the meaning of the words more specific (low uncertainty), whereas usage of third person pronouns might indicate higher uncertainty. Here we used a simple attribute *Modal Verbs*, which is the ratio of modal verbs to the total number of verbs in a page.

Affect: For computing the affect of pages we have utilized SentiWordNet [8], a freely available lexical resource (integrated in *Corleone*) in which each synset s of WordNet[2] is associated with three numerical scores, namely, *Pos(s)*, *Neg(s)* and *Obj(s)*, which describe how 'positive', 'negative' and 'objective' the terms contained in synset s are. In particular, we defined two attributes *PosSent* and *NegSent* for computing the affect of text $T = t_1 \ldots t_n$, where t_i is the i-th token in the text. They are calculated 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'.

Finally, we utilized an in-house list of positive and negative opinion expressions (around 1200) and calculated *Tonality*, an unweighted affect score, which is the ratio of 'positive' tonality expressions to all opinion expressions.

2.2 Features obtained with General Inquirer

General Inquirer's (*GI*) 182 categories were developed for social-science content-analysis applications. The values assigned by *General Inquirer* 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 linguistic feature. A snapshot of category types covered by *GI* is given in Table 1, more details are available at the Web page.[3]

3. EXPERIMENTS

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

Figure 1: Overview of *GI* categories.

This section describes in detail the carried experiments and the data set.

3.1 Datasets

We used two Web spam data sets for our experiments: *WebSpam-Uk2006* [4, 1] and *WebSpam-Uk2007* [1]. These data sets concern two general crawls of the `.uk` Web domain and provide each page's content, links and human-assessed categories of each host (spam, non-spam, borderline, undecided). Note that, at the time of writing, only partial set of labels was available for the 2007 data set.

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[4] running on a cluster of 10 quad core machines was used for all computations. The data set statistics are given in the 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

Table 1: Simple statistics of input data sets (compr.: compressed files, SQF: Hadoop's sequence file).

	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 (compr., GB)	14.10	45.50
HTML SQF (compr., GB)	11.70	37.80
TXT SQF (compr., GB)	2.87	8.24

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

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

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

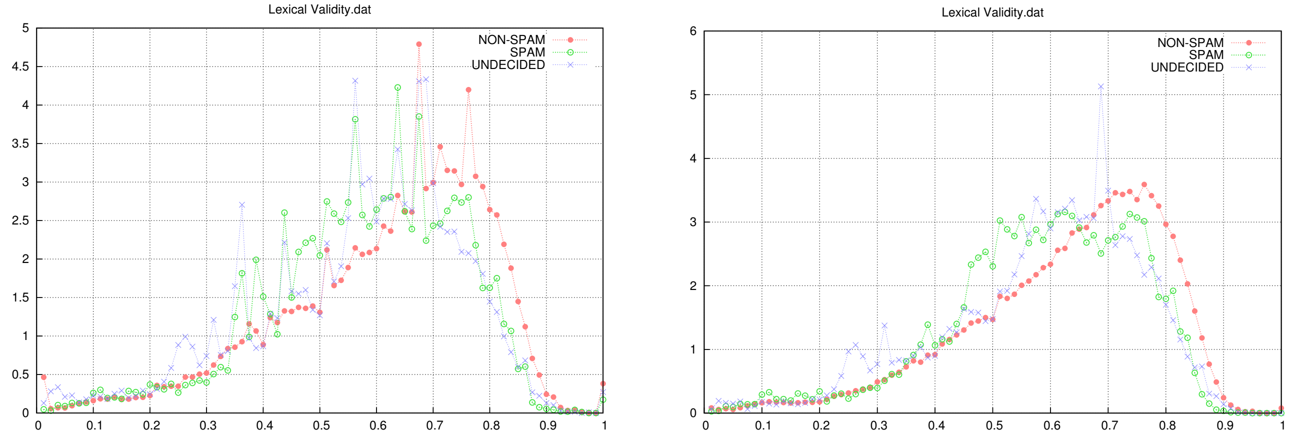

Figure 2: Example of influence of size-filtering on noise-level. Lexical Validity for unfiltered input (left) and mode-1 size-filtered input (right), *Corleone*, *WebSpam-Uk2007*. Size-filtering noticeably removes noise.

for *General Inquirer*) for each page satisfying the length constraints of the appropriate mode. Then, for each attribute's value we created a 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 generating 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.

3.3 General observations

Subsequently, the histograms were studied to identify the most 'discriminative' attributes. Preliminary observations seem to indicate that:

– the histograms generated with the length restrictions are noticeably less noisy than the corresponding histograms computed without such restrictions (mode 0), Figure 2 shows an example of this,

– mode-2 histograms seemed to be a bit less noisy than mode-1 histograms. Due to this observations, we used only the mode-2 attributes for computing the statistics in the next section.

– the borderline histograms are (in general) rather closer to spam than to normal class, but this fact is to be studied more rigorously,

3.4 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:

$$absDist(h) = \sum_{i \in I} |s_i^h - n_i^h|/200 \qquad (1)$$

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:

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

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 the both metrics the higher values indicate better discriminative power of the considered attribute.

Next, for each out of over 1200 histograms we computed both measures and, for each of the 6 settings, we sorted the values non-increasingly 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*'s attributes (according to any distance measure) is the same for both data sets (2006 and 2007) despite the fact that the 2 metrics are quite different. The results are presented in the Tables 2 and 3. Notice, in the Table 2, that some histograms differ by almost 25% of the AUC, which seems to indicate that they can be quite promising as class discriminators. Figure 4 shows an example of a discriminative (*Content diversity*) and non-discriminative (*Capitalized tokens*) attribute computed with the *Corleone* tool. Figure 5 shows the histograms for *Syntactical diversity* based on 4-grams, which seems to be the overall 'winner'.

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 the Table 4). Notice that for some attributes the histograms for spam and non-spam differ on 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 5). There is a significant overlap between the 2 lists of attributes identified by the both, quite different metrics.

Table 2: The most discriminating *Corleone*'s attributes according to the *absDist* metric.

Corleone(absDist)	2007	2006
Passive Voice	0.263	0.273
Syntactical Diversity, 4-Grams	0.255	0.245
Content Diversity	0.234	0.331
Syntactical Diversity, 3-Grams	0.230	0.253
Pronoun Fraction	0.224	0.261
Syntactical Diversity, 2-Grams	0.221	0.232
Lexical Diversity	0.213	0.262
Syntactical Entropy, 2-Grams	0.208	0.179
Text-Like Fraction	0.188	0.184

Table 3: The most discriminating *Corleone*'s attributes according to the *sqDist* metric.

Corleone (sqDist)	2007	2006
Syntactical Diversity, 4-Grams	0.053	0.054
Syntactical Diversity, 3-Grams	0.050	0.067
Syntactical Diversity, 2-Grams	0.037	0.036
Content Diversity	0.032	0.065
Syntactical Entropy, 2-Grams	0.029	0.026
Lexical Diversity	0.026	0.043
Lexical Validity	0.024	0.033
Pronoun Fraction	0.024	0.031
Text-Like Fraction	0.023	0.017

Some other *GI*-generated attributes were identified as among the top-10 by one or another measure or setting. Among them were: *EnlOth*, *WltTot*, *Exch*, *ECON*, *Objects* (see figure 3 for their description). 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 sizes of the full sets are different (*GI* set is over 7 times bigger than the *Corleone* set). Figure 6 demonstrates histograms for the *leftovers* attribute for the 2006 data set.

3.5 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,

Table 4: The most discriminating *General Inquirer's* attributes identified by the *absDist* metric.

GI(absDst)	2007	2006
WltTot	0.287	0.346
WltOth	0.285	0.341
Academ	0.270	0.263
Object	0.255	0.282
EnlTot	0.249	0.247
Econ@	0.228	0.356
SV	0.206	0.26

Table 5: The most discriminating *General Inquirer's* attributes identified by the *sqDist* metric.

GI(sqDist)	2007	2006
leftovers	0.0150	0.0128
EnlOth	0.0085	0.0072
EnlTot	0.0082	0.0118
Object	0.0073	0.0086
text-length	0.0056	0.0048
ECON	0.0038	0.0034
Econ@	0.0038	0.0031
WltTot	0.0038	0.0027
WltOth	0.0037	0.0024

- *WltTot*, *WltOth*: words for pursuit of wealth, such as buying and selling, other words in wealth domain including economic domains and commodities
- *Econ@dat*, *ECON*: words of an economic, commercial, industrial, or business orientation, including roles, collectivities, acts, abstract ideas, and symbols
- *EnlTot*, *EnlOth*: enlightment words (knowledge, insight, and information concerning personal and cultural relations), including words that reflect enlightenment gain (through thought, education), enlightment loss (misunderstanding, being misguided), enlightments participants (e.g., words referring to roles in the secular enlightenment sphere), etc.
- *Objects*: words referring to objects, including tools, food, vehicles, buildings, tools of communication, natural objects other than people and animals (e.g., plants, minerals), body parts, etc.
- *Leftovers*: encompasses several 'Lasswell dictionary' attributes not associated with any other 'large' categories in *GI*, and includes words of (non-)accomplishment, words of transaction or exchange, words referring to means and utility or lack thereof, words of desired or undesired ends or goals, words denoting actors, nations and emotions such as nihilism, disappointment and futility.

Figure 3: Description of most discriminating *General Inquirer's* attributes.

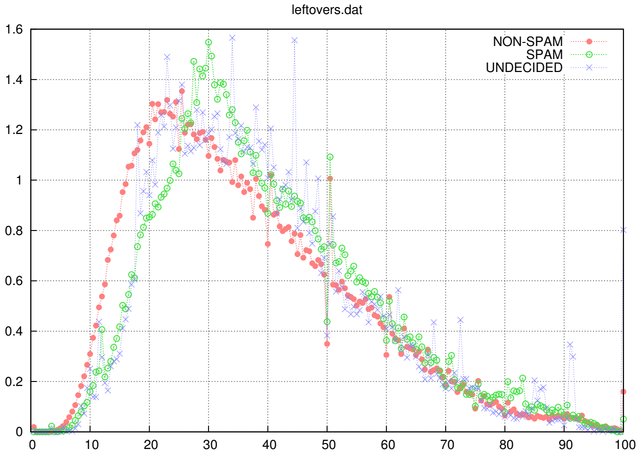

Figure 6: Leftovers attribute, *General Inquirer*, mode-1 filtered set of pages from *WebSpam-Uk2006*.

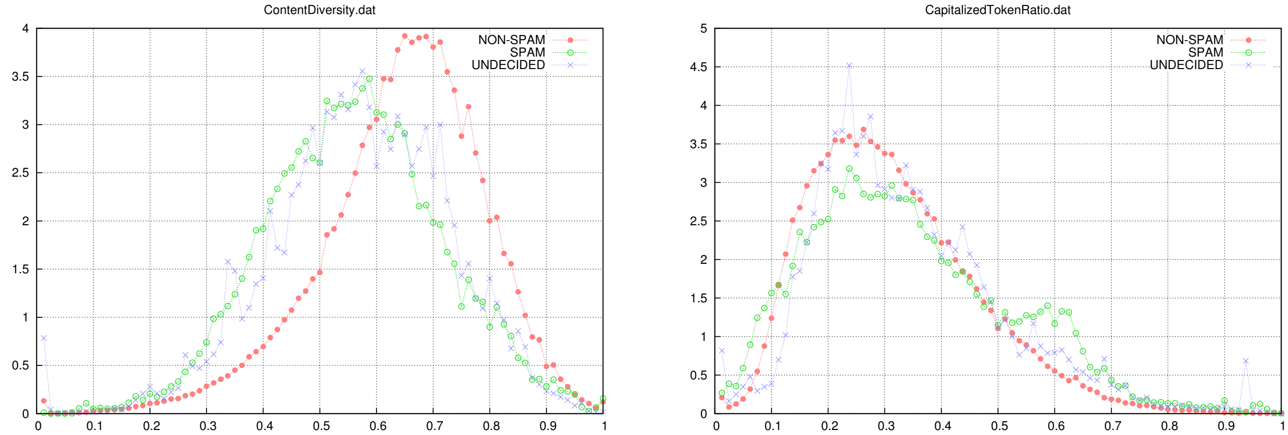

Figure 4: Content diversity (left) and Capitalized tokens (right), *Corleone* processor, mode-1 filtered set of pages from *WebSpam-Uk2006*.

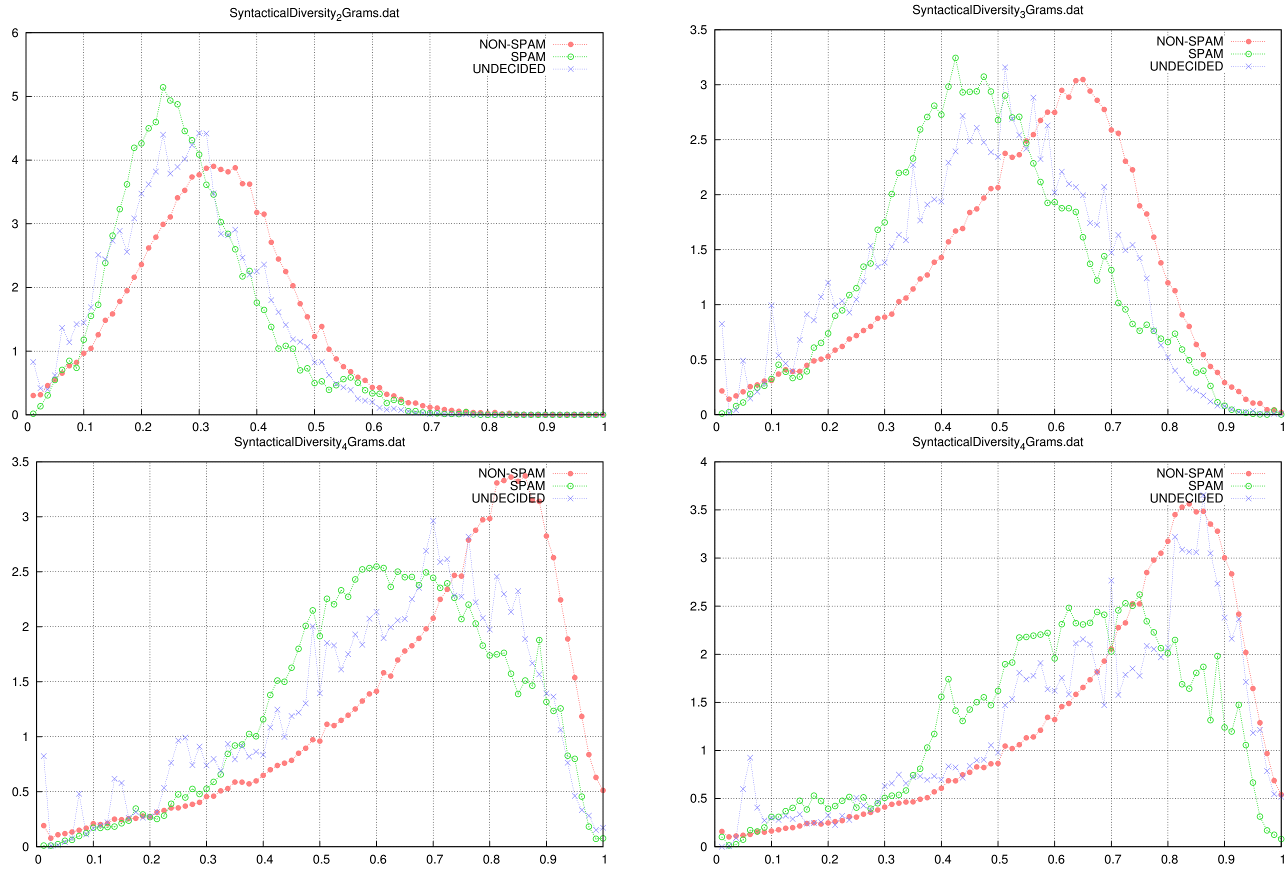

Figure 5: Syntactical diversity for various inputs and n-grams. Top row, left-to-right: 2006 data set, mode-1 filtered input, 2-grams (left), 3-grams (right). Bottom row, left-to-right: 2006 data set, mode-1 filtered input, 4-grams (left), 2007 data set, 4-grams (right). Charts have different scale along the Y axis to visualize distribution shape similarities. Note increasing distribution skew for the normal class with larger n-grams.

economy, business, industry, non-human objects (e.g., food, tools, etc.) and enlightment.

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 [16], 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 computation and preliminary experimentation on over 200 linguistic-based attributes on 2 publicly available Web-spam-reference corpora. To the best of the authors' knowledge, such attributes have not been previously studied in the context of Web spam detection. We discuss the general properties of over 1200 analyzed histograms. With 2 introduced distribution difference metrics, we identify the most promising attributes w.r.t. to the extent of difference in their distributions in the spam and non-spam classes. The list of promising attributes seems to be stable across a couple of different settings, however their real usefulness in the classification task is to be studied in further work. All the computed and studied attributes and histograms are publicly available to the research community.

Acknowledgments. We used Web spam reference corpora *WebSpam-Uk2006* and *WebSpam-Uk2007* [1]. We would like to thank Aleksander Wawer for directing our attention to *General Inquirer* package. We also thank all the contributors of the Hadoop project for their perfect work. We used the HPC cluster at the Institute of Computing Science, Poznan University of Technology for computations. The work presented in this paper was supported by the Polish-Japanese Institute of Information Technology research grant ST/SI/06/2007 'Automatic methods for search engine spam detection'.

5. REFERENCES

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

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

[3] A. Benczúr, I. Bíró, K. Csalogány, and T. Sarlós. Web spam detection via commercial intent analysis. In *AIRWeb '07: Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, pages 89–92, New York, NY, USA, 2007. ACM.

[4] C. Castillo, D. Donato, L. Becchetti, P. Boldi, S. Leonardi, M. Santini, and S. Vigna. A reference collection for web spam. *SIGIR Forum*, 40(2):11–24, December 2006.

[5] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri. Know your neighbors: web spam detection using the web topology. In W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, and N. Kando, editors, *SIGIR*, pages 423–430. ACM, 2007.

[6] I. Drost and T. Scheffer. Thwarting the nigritude ultramarine: learning to identify link spam. In *Proceedings of the 16th European Conference on Machine Learning (ECML)*, volume 3720 of *Lecture Notes in Artificial Intelligence*, pages 233–243, Porto, Portugal, 2005.

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

[8] A. Esuli and F. Sebastiani. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC-06, 5th Conference on Language Resources and Evaluation*, pages 417–422, Genova, IT, 2006.

[9] D. Fetterly, M. Manasse, and M. Najork. Spam, damn spam, and statistics: using statistical analysis to locate spam web pages. In *WebDB '04: Proceedings of the 7th International Workshop on the Web and Databases*, pages 1–6, New York, NY, USA, 2004. ACM.

[10] D. Fetterly, M. Manasse, and M. Najork. Detecting phrase-level duplication on the world wide web. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 170–177, New York, NY, USA, 2005. ACM.

[11] Jakub Piskorski. Corleone - Core Linguistic Entity Extraction. Technical Report (in progress). Joint Research Center of the European Commission, 2008.

[12] G. Mishne, D. Carmel, and R. Lempel. Blocking blog spam with language model disagreement. In *Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, May 2005.

[13] 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.

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

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

[16] 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.