# Spam filtering methods

## Naive Bayesian classification

### Theory

Each message is represented by a vector Following Sahami et al., we use binary attributes: if some characteristic represented by xi is present in the message; otherwise In our experiments, attributes correspond to words, i.e. each attribute shows if a particular word (e.g. “adult”) is present. To select among all possible attributes, we follow Sahami et al. and compute the mutual information (MI) of each candidate attribute X with the category-denoting variable C:

.

The probabilities P(X|C) are practically impossible to estimate directly (the possible values of X are too many, and there are data-sparseness problems). The Naive Bayesian classifier makes the simplifying assumption that X are conditionally independent given the category C . Then:

,

where P(Xi | C) can be easily estimated as relative frequencies from the training corpus;

P(C) - relative frequencies from the training corpus.

### Previous results

Table 1 shows the results of the study by Sahami et al. according to the spam classification. The study used 500 attributes with a threshold of 0.999. The results show the percentage of spam, the accuracy of spam classification and the completeness of spam classification for three different sets of attributes. All sets of attributes have high accuracy and completeness of spam classification.

Table 1 – Results of Sahami et al. (500 attributes, threshold = 0.999, 999 = λ)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Attributes | Total Messages | Testing Messages | % Spam | Spam Precision | Spam Recall |
| words only | 1789 | 251 | 88.2% | 97.1% | 94.3% |
| words+phrases | 1789 | 251 | 88.2% | 97.6% | 94.3% |
| phrases+non-textual | 1789 | 251 | 88.2% | 100.0% | 98.3% |

### Cost-sensitive evaluation measures

In classification tasks, two commonly used evaluation measures are accuracy (Acc) (1) and error rate ( Err = 1- Acc) (2). In our case:

|  |  |
| --- | --- |
|  | (1) |
|  | (2) |

Table 2 – Results on Ling-Spam for best no. of attributes (2893 total messages, 16.6% spam, 10-fold cross validation, attributes ranging from 50 to 700 by a step of 50)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Filter Configuration | λ | No. of attrib. | Spam Recall | Spam Precision | Weighted Accuracy | Baseline W. Acc. | TCR |
| bare | 1 | 50 | 82.10% | 96.85% | 99.02% | 76.11% | 3.42 |
| stop-list | 1 | 50 | 83.70% | 96.926% | 96.85% | 83.374 | 6.71 |
| Bare | 9 | 100 | 82.60% | 93.374% | 97.064% | 81.10% | 5.66 |
| stop-list | 9 | 100 | 88.10% | 96.85% | 95.37% | 78.41% | 4.32 |
| Bare | 999 | 200 | 89.10% | 99.02% | 99.02% | 76.12% | 2.11 |
| Bare | 999 | 200 | 89.30% | 97.064% | 96.926% | 79.361% | 6.71 |

At all three λ values, the highest TCR scores were obtained with the lemmatize enabled. Figure 1 and Figure 2 show that, the stop-list had an additional positive effect for λ=1 and λ=9, but not for λ=999, what is shown in Figure 3. The differences, however, are not always statistically significant. For λ=1, paired single-tailed t-tests on WAcc between all filter configurations of Table 2 confirm only that configuration (b) and (d) are better than (a) at p<0.05.

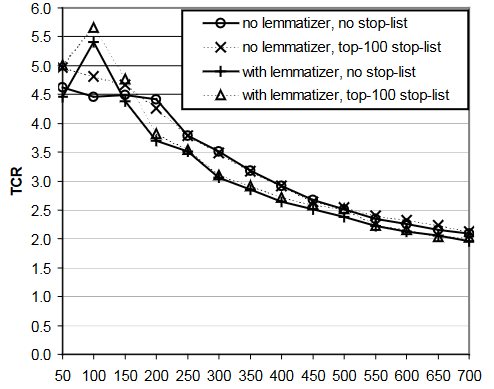


Figure 1 – TCR at t =0.01 (λ=1)

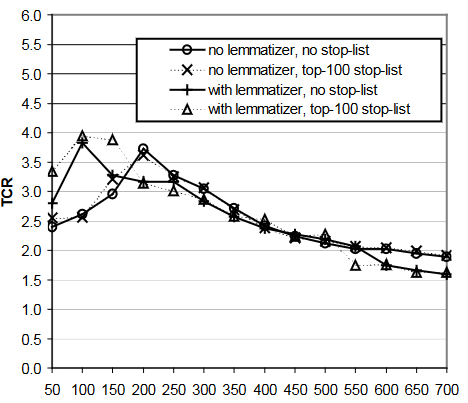


Figure 2 – TCR at t=0.9 (λ=9)

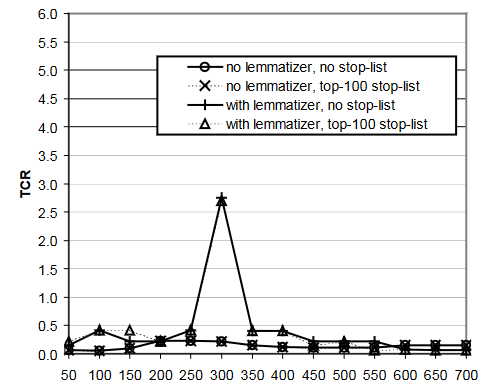


Figure 3 – TCR at t=0.999 (λ=999)

## Link analysis

### Review

The Web is both an excellent medium for sharing information and an attractive platform for delivering products and services. This platform is, to some extent, mediated by search engines in order to meet the needs of users seeking information. Search engines are the “dragons” that keep a valuable treasure: information [Gori and Witten 2005]. Given the vast amount of information available on the Web, it is customary to answer queries with only a small set of results (typically 10 or 20 pages at most). Search engines must then rank Web pages, in order to create a short list of high-quality results for users. On the other side, the Web contains numerous profit-seeking ventures that are attracted by the prospect of reaching millions of users at a very low cost. A large fraction of the visits to a Web site originates from search engines and most of the users click on the first few results in a search engine. Therefore, there is an economic incentive for manipulating search engines’ listings by creating pages that score high independently of their real merit. The term spam has been commonly used in the Internet era to refer to unsolicited (and possibly commercial) bulk messages. The most common form of electronic spam is e-mailing spam, but in practice each new communication medium has created a new opportunity for sending unsolicited messages. These days there are many types of electronic spam, including spam by instant messaging (spam), spam by internet telephony (spit), spam by mobile phone, by fax, etc. The Web is not absent from this list, but as the request response paradigm of the HTTP protocol makes it impossible for spammers to actually send pages directly to the users, spammers try to deceive search engines and thus break the trust that search engines establish with their users.

### General Algorithm

1. Define a set of calculation objects;
2. Expand the page collection;
3. Convert to an undirected bipartite graph;
4. Link Distribution;
5. Setting edges in the node relationship graph;
6. Calculation of the authoritative weight.

### Advantages and disadvantages of link analysis for spam filtering

Advantages:

* effective in identifying and blocking spam emails that contain malicious links or deceptive URLs;
* helps in detecting patterns and relationships between different spam emails, allowing for more accurate and targeted filtering;
* can be integrated with other spam filtering techniques to improve overall detection rates;
* provides insights into the behavior and tactics of spammers, allowing for better adaptive filtering strategies;
* helps in reducing the burden on users by automatically filtering out most malicious or unwanted emails.

Disadvantages:

* vulnerable to evasion techniques used by spammers, such as obfuscation or URL redirection;
* can produce false positives if legitimate emails containing harmless links are incorrectly flagged as spam;
* requires continuous updating of link databases and analysis algorithms to keep up with evolving spam tactics;
* may not be effective against certain types of spam emails that do not contain malicious links;
* limited in its ability to detect non-link-based spam content, such as phishing emails or image-based spam.

Сonclusion

Spam filtering is a process of identifying and filtering out unwanted or unsolicited emails, messages, or comments that are sent in bulk to multiple recipients. It helps to reduce the amount of spam that users receive in their inbox and protects them from potentially harmful content such as phishing scams or malware. Spam filters use various techniques such as content analysis, blacklists, whitelists, and artificial intelligence algorithms to automatically detect and block spam messages before they reach the recipient's inbox.