Feature engineering for mobile (SMS) spam filtering

Conference Paper · January 2007

DOI: 10.1145/1277741.1277951 · Source: DBLP

3 authors, including:

Jose Maria Gomez Hidalgo
TIBCO
78 PUBLICATIONS **1,157** CITATIONS
SEE PROFILE

Enrique Puertas Sanz
European University of Madrid
52 PUBLICATIONS **465** CITATIONS
SEE PROFILE

Feature Engineering for Mobile (SMS) Spam Filtering

Gordon V. Cormack
School of Computer Science
University of Waterloo
Waterloo, Ontario N2L 3G1, CANADA
gvcormac@uwaterloo.ca

José María Gómez Hidalgo
Universidad Europea de Madrid
Villaviciosa de Odón
28670 Madrid, SPAIN
jmgomez@uem.es

Enrique Puertas Sánz
Universidad Europea de Madrid
Villaviciosa de Odón
28670 Madrid, SPAIN
enrique.puertas@uem.es

ABSTRACT
Mobile spam in an increasing threat that may be addressed using filtering systems like those employed against email spam. We believe that email filtering techniques require some adaptation to reach good levels of performance on SMS spam, especially regarding message representation. In order to test this assumption, we have performed experiments on SMS filtering using top performing email spam filters on mobile spam messages using a suitable feature representation, with results supporting our hypothesis.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and retrieval – *information filtering*.

H.3.4 [**Information Storage and Retrieval**]: Systems and Software – *performance evaluation (efficiency and effectiveness).*

General Terms
Performance, Experimentation, Security, Standardization.

Keywords
Spam filtering, Mobile Spam, SMS, ROC Analysis, TREC.

1. INTRODUCTION
SMS spam is now prevalent in Singapore and Japan and will undoubtedly spread throughout the world. As mobile devices increase in computational power, and sophisticated and powerful systems can be connected to mobile phone networks, it is wise to test which technical measures against email spam can be transferred to SMS spam. We report here our work on making current email spam filters effective on mobile spam.

It is not clear that current spam filters should perform well on mobile spam. SMS messages are shorter then email; they lack structured fields, and their text is rife with abbreviations and idioms. We have performed a series of experiments on SMS spam filtering, using high performance email spam filters, following TREC-like procedures [2], and focusing on feature definition. The results of our experiments show that feature engineering is more critical for mobile spam filtering than for email filtering.

2. DATA COLLECTION & PROCESSING
In order to test filters on SMS spam, a collection of spam and ham (not spam) messages must be collected. We use variations of the collections described in [3]:

- A collection of English SMS messages, including 1002 legitimate messages randomly extracted from the NUS SMS Corpus and the Jon Stevenson Corpus, and 82 SMS spam messages collected from the Grumbletext mobile spam site.
- A collection of Spanish SMS messages, donated by Vodafone, including 1157 ham messages obtained in a joke contest, and 199 spam messages reported by users.

Most of the filters used in our experiments do not require any explicit processing of the messages to be run on them. However, machine learning methods require explicit representation as feature vectors. We have followed [4] and provided them with a vector representation using the following features:

- Words – sequences of alpha-numeric characters in the message text. We consider that any non-alphanumeric character is a separator.
- Lowercased words – lowercased words in the text message, according to the definition of word above.
- Character bi-grams and tri-grams – sequences of 2 or 3 characters included in any lowercased word. This attributes try to capture morphological variance and regularities in a language-independent way.
- Word bi-grams – sequences of 2 words in a window of 5 words preceding the current word. This is a version of the Orthogonal Sparse Bigrams method mentioned below.

3. SPAM FILTERS
We have selected a number of high performing filters according to TREC [2] evaluations:

- Bogofilter – a popular open-source Bayesian spam filter that performed well at TREC;
- DMC – Dynamic Markov Compression – a adaptive method based on the DMC compression method;
- LR – TR-IRLS – an established open-source logistic regression classifier;
- OSBF-Lua [1] – Orthogonal Sparse Bigrams with confidence Factor – a Bayesian classifier enhanced with Orthogonal Sparse Bigrams for feature extraction, and Exponential Differential Document Count for automatic feature selection;
- SVM – SVMlight – an established free-for-scientific use support vector machine classifier.

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.
ACM 978-1-59593-597-7/07/0007.

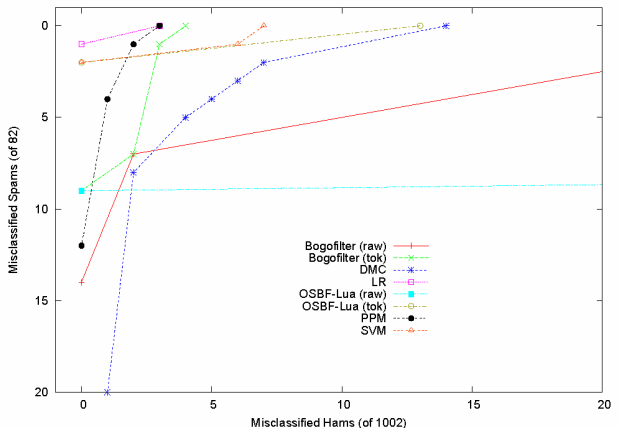

Figure 1. ROC curve – English collection.

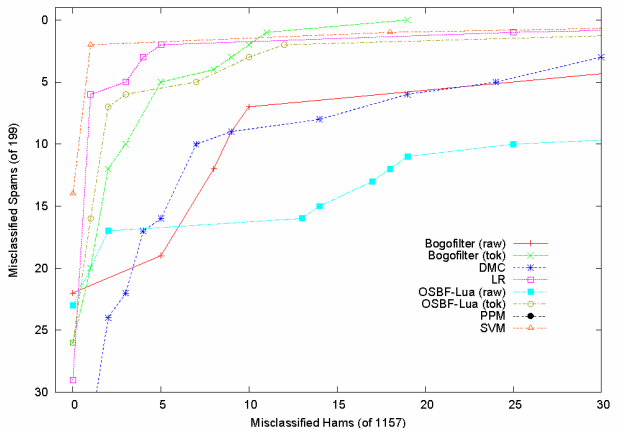

Figure 2. ROC curve – Spanish collection.

These spam filters[1] have performed consistently well on TREC spam filtering evaluations, OSBF-Lua being the top scoring method in TREC 2006. The filters were presented the messages in raw form, but Bogofilter and OSBF-Lua were also fed a textual representation of the feature vectors discussed above: Each feature is represented by a different dummy word, repeated as many times as the feature appears in the original message text.

4. EVALUATION AND DISSCUSION

We evaluated each filter on each corpus using 10-fold cross validation. Following TREC, we plot the tradeoff between ham misclassification and spam misclassification as a receiver operating characteristic (ROC) curve. As a summary measure we report 1-AUC as a percentage where AUC is the area under the ROC curve (normalized so that each axis extends from 0 to 1).

Figures 1 and 2 show the ROC curves for the English and Spanish collections; (1-AUC)% statistics are presented in table 1, along with a 95% confidence interval (the smaller, the better).

Bogofilter and OSBF-Lua perform poorly on the raw messages, but are competitive using our textualized features. DMC and PPM, which are not feature based, perform well without modification. Logistic Regression and SVM perform well on our features. In absolute terms the performance of all filters (except the raw versions of Bogofilter and OSBF-Lua) is comparable to what one might expect for an email corpus of comparable size. We note that the differences in AUC among these six filters are not statistically significant – a larger corpus will be necessary to distinguish them.

The effect of shorter and sparser text is also clear. Bogofilter and OSBF-Lua perform poorly on the raw messages and much better on the textualized feature vectors. OSBF-Lua reports the fewest mistakes, with only five false negatives and no false positives, for the English collection.

As a final conclusion, the differences among all the filters are not clear, so more experiments with a larger dataset are required.

[1] Pointers to descriptions of these spam filters and Machine Learning methods can be found in e.g. [2].

5. FUTURE WORK

Our next steps include building a bigger test collection (from 3 to 10 times bigger) and testing other algorithms and tokenization methods on it, in order to confirm these first results. We are also considering the scenarios for which online evaluation makes sense. Finally, we believe that this work may be an interesting starting point for including a new dataset and task in TREC evaluation.

Table 3. Scores of the filters, (1-AUC)% with confidence intervals, for the Spanish and English collections

Method	English SMS	Spanish SMS
Bogo (raw)	1.308 (0.811 - 2.102)	1.279 (0.782 - 2.084)
Bogo (tok)	0.116 (0.036 - 0.371)	0.191 (0.105 - 0.349)
DMC	0.144 (0.048 - 0.431)	0.166 (0.077 - 0.355)
LR	0.165 (0.015 - 1.715)	0.132 (0.047 - 0.369)
OSBF (raw)	5.721 (3.750 - 8.634)	2.207 (1.463 - 3.318)
OSBF (tok)	0.238 (0.033 - 1.662)	0.268 (0.131 - 0.545)
SVM	0.210 (0.038 - 1.137)	0.110 (0.035 - 0.343)

6. ACKNOWLEDGMENTS

We would like to thank Vodafone for making the Spanish SMS spam collection available to us and for partially funding these experiments.

7. REFERENCES

[1] Assis F., OSBF-Lua - A Text Classification Module for Lua. The Importance of the Training Method, *The Fifteenth Text REtrieval Conference Proceedings (TREC)*, Nov, 2006.

[2] Cormack G.V. and Bratko A., Batch and On-line Spam Filter Evaluation, *CEAS 2006 - Third Conference on Email and Anti-Spam*, Mountain View, July 2006.

[3] Gómez, J.M., Cajigas, G., Puertas Sanz, E., Carrero García, F. Content Based SMS Spam Filtering, *Proceedings of the 2006 ACM Symposium on Document Engineering*, Amsterdam, The Netherlands, ACM Press. Oct., 2006.