Contributions to the Study of SMS Spam Filtering: New Collection and Results

Tiago A. Almeida
School of Electrical and
Computer Engineering
University of Campinas
Campinas, Sao Paulo, Brazil
tiago@dt.fee.unicamp.br

José María Gómez
Hidalgo
R&D Department
Optenet
Las Rozas, Madrid, Spain
jgomez@optenet.com

Akebo Yamakami
School of Electrical and
Computer Engineering
University of Campinas
Campinas, Sao Paulo, Brazil
akebo@dt.fee.unicamp.br

ABSTRACT

The growth of mobile phone users has lead to a dramatic increasing of SMS spam messages. In practice, fighting mobile phone spam is difficult by several factors, including the lower rate of SMS that has allowed many users and service providers to ignore the issue, and the limited availability of mobile phone spam-filtering software. On the other hand, in academic settings, a major handicap is the scarcity of public SMS spam datasets, that are sorely needed for validation and comparison of different classifiers. Moreover, as SMS messages are fairly short, content-based spam filters may have their performance degraded. In this paper, we offer a new real, public and non-encoded SMS spam collection that is the largest one as far as we know. Moreover, we compare the performance achieved by several established machine learning methods. The results indicate that Support Vector Machine outperforms other evaluated classifiers and, hence, it can be used as a good baseline for further comparison.

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, classification

1. INTRODUCTION

Short Message Service (SMS) is the text communication service component of phone, web or mobile communication systems, using standardized communications protocols that allow the exchange of short text messages between fixed line or mobile phone devices. According to the International Telecommunication Union (ITU)[1], SMS has become a massive commercial industry, worth over 81 billion dollars globally as of 2006.

The downside is that cell phones are becoming the latest target of electronic junk mail, with a growing number of marketers using text messages to target subscribers. SMS spam (sometimes also called mobile phone spam) is any junk message delivered to a mobile phone as text messaging. Although this practice is rare in North America, it has been very common in some parts of Asia.

Apparently, SMS spam is not as cost-prohibitive to spammers as it used to be, as the popularity of SMS has led to messaging charges dropping below US$ 0.001 in markets like China, and even free of charge in others. According to Cloudmark stats[2], the amount of mobile phone spam varies widely from region to region. For instance, in North America, much less than 1% of SMS messages were spam in 2010, while in parts of Asia up to 30% of messages were represented by spam.

The main problem with SMS spam is that it is not only annoying, but it can also be expensive since some people pay to receive text messages. Moreover, there is a limited availability of mobile phone spam-filtering software. Other concern is that important legitimate messages as of emergency nature could be blocked. Nonetheless, many providers offer their subscribers means for mitigating unsolicited SMS messages.

In the same way that carriers are facing many problems in dealing with SMS spam, academic researchers in this field are also experiencing difficulties. For instance, the lack of real and public databases can compromise the evaluation of different approaches. So, although there has been significant effort to generate public benchmark datasets for anti-spam filtering, unlike email spam, which has available a large variety of datasets, the mobile spam filtering still has very few corpora usually of small size. Other concern is that established email spam filters may have their performance seriously degraded when directly employed to dealing with mobile spam, since the standard SMS messaging is limited

---

[1]See http://www.itu.int/dms_pub/itu-s/opb/pol/S-POL-IR.DL-2-2006-R1-SUM-PDF-E.pdf.
[2]See http://www.cloudmark.com/en/article/mobile-operators-brace-for-global-surge-in-...mobile-messaging-abuse

to 140 bytes, which translates to 160 characters of the English alphabet. Moreover, their text is rife with idioms and abbreviations.

To fill this important gap, in this paper, we make available a new real, public and non-encoded SMS spam corpus that is the largest one as far as we know. Moreover, we compare the performance achieved by several established machine learning methods in order to provide good baseline results for further comparison.

The remainder of this paper is organized as follows. Section 2 offers details about the newly-created SMS Spam Collection. In Section 3, we present a comprehensive performance evaluation for comparing several established machine learning approaches. Finally, Section 4 presents the main conclusions and outlines for future works.

2. THE NEW SMS SPAM COLLECTION

Reliable data are essential in any scientific research. The absence of representative data can seriously impact the processes of evaluation and comparison of methods. In this way, areas of more recent studies are generally affected by the lack of public available data.

Regarding studies of mobile spam filtering, although there are few databases of legitimate SMS messages available in the Internet, it is very hard to find real samples of mobile phone spam. Thus, to create the corpus for the purposes of this work, we use data derived from several sources.

A collection of 425 SMS spam messages was manually extracted from the Grumbletext Web site. This is a UK forum in which cell phone users make public claims about SMS spam messages, most of them without reporting the very spam message received. The identification of the text of spam messages in the claims is a very hard and time-consuming task, and it involved carefully scanning hundreds of web pages. The Grumbletext Web site is: `http://www.grumbletext.co.uk/`.

We have also included in our corpus a subset of 3,375 SMS randomly chosen ham messages of the NUS SMS Corpus, which is a dataset of about 10,000 legitimate messages collected for research at the Department of Computer Science at the National University of Singapore. These messages were collected from volunteers who were made aware that their contributions were going to be made publicly available. The NUS SMS Corpus is available at: `http://www.comp.nus.edu.sg/~rpnlpir/downloads/corpora/smsCorpus/`.

We have added legitimate samples by inserting 450 SMS messages collected from Caroline Tag's PhD Thesis available at `http://etheses.bham.ac.uk/253/1/Tagg09PhD.pdf`.

Finally, we have incorporated the SMS Spam Corpus v.0.1 Big. It has 1,002 SMS ham messages and 322 spam messages and is public available at: `http://www.esp.uem.es/jmgomez/smsspamcorpus/`. This corpus has been used in the following academic research efforts: [6], [7], and [14]. The sources used in this corpus are also the Grumbletext Web site and the NUS SMS Corpus.

The created corpus is composed by just one text file, where each line has the correct class followed by the raw message.

The SMS Spam Collection is public available at `http://www.dt.fee.unicamp.br/~tiago/smsspamcollection`.

The new collection is composed by 4,827 legitimate messages and 747 mobile spam messages, a total of 5,574 short messages. To the best of our knowledge, it is the largest

available SMS spam corpus that currently exists. Table 1 shows the basic statistics of the created database.

Table 1: Basic statistics

Msg	Amount	%
Hams	4,827	86.60
Spams	747	13.40
Total	5,574	100.00

Table 2 presents the statistics related to the tokens extracted from the corpus. Note that, the proposed dataset has a total of 81,175 tokens and mobile phone spam has in average ten tokens more than legitimate messages.

Table 2: Token statistics

Hams	63,632
Spams	17,543
Total	81,175
Avg per Msg	14.56
Avg in Hams	13.18
Avg in Spams	23.48

2.1 Message Duplicates Analysis

As the newly collected messages in the SMS Spam Collection have been augmented with a previously existing database built using roughly the same sources (GrumbleText, NUS SMS Corpus), it is sensible to check if there are some duplicates coming from both databases.

To address this issue, we have performed a duplicates analysis based on plagiarism detection techniques [9]. In [9], a number of techniques for plagiarism are presented, and those based on "String-of-Text", and implemented by the tool WCopyfind[3], can be considered as reasonable baseline for the purpose of detecting near-duplicated messages in our collection. The "String-of-Text" methods involve scanning suspect texts for approximately matching character sequences. Texts are compared searching for N-grams for relatively big sizes (*e.g.* $N = 6$), with additional parameters (length of match in number of characters, etc.). We have simplified this method to N-gram matches after text normalization involving replacing all token separators by white spaces, lowercasing all characters, and replacing digits by the character 'N' (to preserve phone numbers structure).

We searched for near-duplicates within three subcollections: the previously existing SMS Spam Corpus v.0.1 Big (**INIT**), the additional messages from Grumbletext, the NUS SMS Corpus, and Tag's PhD Thesis (**ADD**), and the actual new SMS Spam Collection (**FINAL**). In order to assess the overlap between both collections, we have compared each pair of messages within each subcollection and in common between both subcollections, stored all N size matches (N-grams with $N = 5$, 6, and 10), and sorted the N-grams according to their frequencies.

We have found 5-grams already presented in the **INIT** and the **ADD** collections do not collapse to greatly increase their frequencies, and they typically correspond to templates often presented in cell phones, and used in legitimate messages (*e.g.* "sorry i ll call later"). The 5-grams that co-occur in **INIT** and **ADD**, so they get their frequencies increased

---

[3]See: `http://plagiarism.phys.virginia.edu`

in **FINAL**, are new instances of spam probably sent by the same organization. In 6-grams results, we have found that there are not significant near-duplicates except for those already presented in each subcollection. Moreover, the results achieved with 10-grams are very similar to the 5- and 6-grams ones.

In consequence, we believe it is safe to say that merging the subcollections, although they have roughly the same sources, does not lead to near-duplicates that may ease the task of detecting SMS spam.

3. EXPERIMENTS

We have tested several well-known machine learning methods in the task of automatic spam filtering using the created SMS Spam Collection. The main goal of this performance evaluation is to provide good baseline results for further comparison, since established email spam filters may have their performance seriously impacted when employed to classify short messages. In addition, mobile phone messages often have a lot of abbreviations and idioms that may also affect the filters accuracy.

We consider in this work two different tokenizers:

1. tok1: tokens start with a printable character, followed by any number of alphanumeric characters, excluding dots, commas and colons from the middle of the pattern. With this pattern, domain names and mail addresses will be split at dots, so the classifier can recognize a domain even if subdomains vary [16].

2. tok2: any sequence of characters separated by blanks, tabs, returns, dots, commas, colons and dashes are considered as tokens. This simple tokenizer intends to preserve other symbols that may help to separate spam and legitimate messages.

In addition, we did not perform language-specific preprocessing techniques such as stop word removal or word stemming, since other researchers found that such techniques tend to hurt spam-filtering accuracy [5, 17].

The list of all evaluated classifiers are presented in Table 3[4].

Given that the corpus is biased to the ham class, an obvious baseline is the trivial rejector (TR) for the spam class.

As the most of the tokens with the highest Information Gain score often occur in the spam class, it is sensible to expect that messages may get automatically segregated into two classes on the basis of those tokens. In consequence, we provide an additional baseline in the form of the results of the Expectation-Maximization (EM) clustering algorithm [8], over a vector representation based on the tokenizer tok2. EM is an iterative soft clusterer that estimates cluster densities. Basically, cluster membership is a hidden latent variable that the maximum likelihood EM method estimates.

In our experiments, we have set up a maximum of 20 iterations and used the rest of the default values for EM parameters in WEKA.

---

[4]Some of the implementations of the described classifiers are provided by the Machine Learning library WEKA, available at `http://www.cs.waikato.ac.nz/ml/weka/`. The algorithms have been used with their default parameters except when otherwise is specified.

Table 3: Evaluated classifiers

Classifiers
Basic Naïve Bayes (NB) – Basic NB [2]
Multinomial term frequency NB – MN TF NB [2]
Multinomial Boolean NB – MN Bool NB [2]
Multivariate Bernoulli NB – Bern NB [2]
Boolean NB – Bool NB [2]
Multivariate Gauss NB – Gauss NB [2]
Flexible Bayes – Flex NB [2]
Boosted NB [12]
Linear Support Vector Machine – SVM [10, 13]
Minimum Description Length – MDL [4]
K-Nearest Neighbors – KNN [1, 14] (K = 1, 3 or 5)
C4.5 [15, 14]
Boosted C4.5 [14]
PART [11, 14]

3.1 Results

We carried out this study using the following experiment protocol. We divided the corpus in two parts: the first 30% of the messages were separated for training (1,674 messages) and the remainder ones for testing (3,900 messages). As all the messages are fairly short, we did not use any kind of method to reduce the dimensionality of the training space, *e.g.*, terms selection techniques.

To compare the results we employed the following well-known performance measures: Spam Caught ($SC\%$), Blocked Hams ($BH\%$), Accuracy ($Acc\%$), and Matthews Correlation Coefficient (MCC) [3].

Table 4 presents the best fifteen results achieved by each evaluated classifier and tokenizer. Note that the results are sorted in descending order of MCC.

Table 4: The fifteen best results achieved by combinations of classifiers + tokenizers and the baselines Expectation-Maximization (EM) and trivial rejection (TR)

Classifier	$SC\%$	$BH\%$	$Acc\%$	MCC
SVM + tok1	83.10	0.18	97.64	0.893
Boosted NB + tok2	84.48	0.53	97.50	0.887
Boosted C4.5 + tok2	82.91	0.29	97.50	0.887
PART + tok2	82.91	0.29	97.50	0.887
MDL + tok1	75.44	0.35	96.26	0.826
C4.5 + tok2	75.25	2.03	95.00	0.770
Bern NB + tok1	54.03	0.00	94.00	0.711
MN TF NB + tok1	52.06	0.00	93.74	0.697
MN Bool NB + tok1	51.87	0.00	93.72	0.695
1NN + tok2	43.81	0.00	92.70	0.636
Basic NB + tok1	48.53	1.42	92.05	0.600
Gauss NB + tok1	47.54	1.39	91.95	0.594
Flex NB + tok1	47.35	2.77	90.72	0.536
Boolean NB + tok1	98.04	26.01	77.13	0.507
3NN + tok2	23.77	0.00	90.10	0.462
EM + tok2	17.09	4.18	85.54	0.185
TR	0.00	0.00	86.95	–

It is notable that the linear SVM achieved the best results and outperformed the other evaluated methods. It caught 83.10% of all spams with the cost of blocking only 0.18%

of legitimate messages, acquiring an accuracy rate higher than 97.5%. It is a remarkable performance considering the EM and TR baselines and the high difficulty of classifying mobile phone messages. However, the results also indicate that the best four algorithms achieved similar performance with no statistical difference. All of them accomplished an accuracy rate superior than 97%, that can be considered as a very good baseline in a such context.

It is important to point out that MDL and C4.5 techniques also achieved good results since they found a good balance between false and true positive rates. On the other hand, the remainder evaluated approaches had an unsatisfying performance. Note that, although the most of them have obtained accuracy rate superior than 90%, they have correctly filtered about only 50% of spams or even less.

Therefore, based on the achieved results, we can certainly conclude that the linear SVM offers the best baseline performance for further comparison.

4. CONCLUSIONS

The task of automatic filtering SMS spam still is a real challenge nowadays. There are three main problems hindering the development of algorithms in this specific field of research: the lack of public and real datasets, the low number of features that can be extracted per message, and the fact that the text is rife with idioms and abbreviations.

To fill some of those gaps, in this paper we presented a new mobile phone spam collection that is the largest one as far as we know. Besides being large, it is also publicly available and composed by only non-encoded and real messages.

Moreover, we offered statistics relating to the proposed corpus, as tokens frequencies and since the corpus is composed by subsets of messages extracted from the same sources, we also presented a study regarding the message duplicates and the found results indicate that the proposed collection is reliable because there are no more duplicates than those ones already presented within the used subsets.

Finally, we compared the performance achieved by several established machine learning methods and the results indicate that SVM outperforms other classifiers and, hence, it can be used as a good baseline for further comparison.

Future work should consider to use different strategies to increase the dimensionality of the feature space. Well-known techniques, such as orthogonal sparse bigrams (OSB), 2-grams, 3-grams, among many others could be employed with the standard tokenizers to produce a larger number of tokens and patterns which can assist the classifier to separate ham messages from spams.

5. ACKNOWLEDGMENTS

6. REFERENCES

[1] D. Aha and D. Kibler. Instance-based learning algorithms. *Machine Learning*, 6:37–66, 1991.

[2] T. A. Almeida, J. Almeida, and A. Yamakami. Spam Filtering: How the Dimensionality Reduction Affects the Accuracy of Naive Bayes Classifiers . *JISA*, 1(3):183–200, 2011.

[3] T. A. Almeida, A. Yamakami, and J. Almeida. Evaluation of Approaches for Dimensionality Reduction Applied with Naive Bayes Anti-Spam Filters. In *Proc. of the 8th IEEE ICMLA*, pages 517–522, Miami, FL, USA, 2009.

[4] T. A. Almeida, A. Yamakami, and J. Almeida. Filtering Spams using the Minimum Description Length Principle. In *Proc. of the 25th ACM SAC*, pages 1856–1860, Sierre, Switzerland, 2010.

[5] G. Cormack. Email Spam Filtering: A Systematic Review. *Foundations and Trends in Information Retrieval*, 1(4):335–455, 2008.

[6] G. V. Cormack, J. M. Gómez Hidalgo, and E. Puertas Sanz. Feature Engineering for Mobile (SMS) Spam Filtering. In *Proc. of the 30th ACM SIGIR*, pages 871–872, New York, NY, USA, 2007.

[7] G. V. Cormack, J. M. Gómez Hidalgo, and E. Puertas Sanz. Spam Filtering for Short Messages. In *Proc. of the 16th ACM CIKM*, pages 313–320, Lisbon, Portugal, 2007.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society.*, 39(1):1–38, 1977.

[9] H. Dreher. Automatic Conceptual Analysis for Plagiarism Detection. *Issues in Informing Science and Information Technology*, 4:601–628, 2007.

[10] G. Forman, M. Scholz, and S. Rajaram. Feature Shaping for Linear SVM Classifiers. In *Proc. of the 15th ACM SIGKDD*, pages 299–308, 2009.

[11] E. Frank and I. H. Witten. Generating Accurate Rule Sets Without Global Optimization. In *Proc. of the 15th ICML*, pages 144–151, Madison, WI, USA, 1998.

[12] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *Proc. of the 13th ICML*, pages 148–156, San Francisco, 1996. Morgan Kaufmann.

[13] J. M. Gómez Hidalgo. Evaluating Cost-Sensitive Unsolicited Bulk Email Categorization. In *Proc. of the 17th ACM SAC*, pages 615–620, Madrid, Spain, 2002.

[14] J. M. Gómez Hidalgo, G. Cajigas Bringas, E. Puertas Sanz, and F. Carrero García. Content Based SMS Spam Filtering. In *Proc. of the 2006 ACM DocEng*, pages 107–114, Amsterdam, The Netherlands, 2006.

[15] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., 1993.

[16] C. Siefkes, F. Assis, S. Chhabra, and W. Yerazunis. Combining Winnow and Orthogonal Sparse Bigrams for Incremental Spam Filtering. In *Proc. of the 8th ECML PKDD*, pages 410–421, Pisa, Italy, 2004.

[17] L. Zhang, J. Zhu, and T. Yao. An Evaluation of Statistical Spam Filtering Techniques. *ACM TALIP*, 3(4):243–269, 2004.