A Kosher Source of Ham

Nathan Friess*
Lyryx Learning, Inc.
210 - 1422 Kensington Road NW
Calgary, Alberta, Canada T2N 3P9

John Aycock†
Department of Computer Science
University of Calgary
2500 University Drive NW
Calgary, Alberta, Canada T2N 1N4

(Appeared in MIT Spam Conference, 2009)

Abstract

Testing content-based anti-spam systems requires a plentiful source of both spam *and* ham. We examine the viability of Usenet postings as a ham source. While Usenet postings have been used before for this purpose, we refine the idea and show empirically that it is the text of Usenet replies that provides the best cut of ham.

1 Introduction

Measuring the efficacy of spam filters and comparing different spam filters is an important task. Filters cannot be methodically improved without measurement; informed choices cannot be made without comparison. However, many spam filters in use today examine the contents of messages as part of their filtering algorithms. Proper testing and measurement of these filters thus requires both spam content and ham content.

For spam, gathering email messages for a corpus is relatively simple. There are numerous existing public corpora, and moreover it is possible to gather new messages using a spam trap. Assuming that the trap email addresses are not confused for real people's addresses, all of the email sent to the spam trap addresses can be immediately flagged as spam even without using a spam classifier.

Gathering ham messages is not so straightforward due to the personal content of ham. Garcia et al. [2] approached this issue by collecting Usenet posts to build their

*Work done while at the University of Calgary.

†Corresponding author; email aycock@ucalgary.ca.

ham corpus. Usenet is a public discussion system where anyone can view or post messages (called articles). The discussions are organized by topic in newsgroups. There are thousands of newsgroups in Usenet, some which are used primarily for exchanging binary files such as images and videos while others are primarily plain-text messages. Because of the open architecture of Usenet, it is a prime target for spammers. A quick trip to Google Groups [3], Google's web-based Usenet interface, reveals the prevalence of spam in most newsgroups.

We also gather ham from Usenet postings, much like Garcia et al., but with one important distinction. We only use Usenet posts that are *replies* to existing discussion threads. The rationale is that Usenet is well known to contain a lot of spam, which we do not want to have in the gathered ham corpus. Therefore, before Usenet postings can be included in a ham corpus they must be vetted, either manually or automatically. Either way, separating out the noise from the actual ham is problematic. Manually verifying thousands of Usenet postings is too time consuming, but using an automated system (such as a spam filter) assumes *a priori* knowledge of Usenet spam and leads to a biased ham corpus. After all, if a spam filter classifies a pool of unknown messages to build a ham corpus, it should be expected that, all else being equal, the same spam filter will then correctly classify all of the ham during the simulation. It should also be noted that both the manual and automated approaches to building a ham corpus will have varying degrees of accuracy, leading to at least some amount of spam leaking into the ham corpus, and vice versa.

2 Experiments

An experienced Usenet user observes that spammers do not seem to reply to existing messages when posting spam; Usenet spam typically appears as new threads which are completely unrelated to existing discussions. In other words, the spam messages lack the proper headers that identify the message as belonging to an existing thread. Similarly, humans rarely reply to spam posts, so seeing a string of replies in a thread is an indicator that the thread as a whole contains ham. Therefore, we hypothesize that by only harvesting replies to Usenet postings, the signal-to-noise ratio will be greatly increased, thus resulting in a ham corpus that is less tainted by spam.

The hypothesis of there being less spam in the set of reply messages was tested by using a spam classifier to compare the proportions of ham and spam in Usenet postings. Two experiments were performed, the first consisting of a list of 38 hand chosen high traffic Usenet groups and the second consisting of a list of 77 Usenet groups from NewsAdmin's top 100 text newsgroups list [5]. The spam classifier used in the experiments was DSPAM [7]. DSPAM is an adaptive classifier, and therefore it must be trained on a corpus before it can be used. For this, the TREC 2005 Spam Track Public Corpus [1] was used. In both the training and classification phases, DSPAM was run with a default configuration under the assumption that DSPAM is tuned for most common situations already. After being trained with the TREC 2005 corpus, all of the messages gathered from the Usenet groups were classified individually by DSPAM, and the verdicts were tallied.

Usenet posts are similar to email messages in that they contain a variable number

```
From: Alice <alice@acme.com>
Newsgroups: rec.food.cooking
Subject: Re: Tasty chicken
Date: Sun, 13 Apr 2008 00:47:32 -0700 (PDT)
Message-ID: <abcd@acme.com>
References: <xyz123@googlegroups.com>

On Apr 12, 7:27 pm, Bob <bob@gmail.com> wrote:
>
>BBQ chicken is the best!
>
>Bob
>

I agree, especially on a hot summer day.

Alice
```

Figure 1: Simple Usenet reply message

of headers and a body [4]. However, some of the headers in Usenet posts are unique to Usenet or have a greater importance than in email. Figure 1 is an example of a Usenet post. Three important message headers are highlighted in bold type. The *Newsgroups* header indicates which group or groups this post belongs to. The *Message-ID* header is a string that uniquely identifies a post, and is generated automatically for every post submitted. While email messages can also contain Message-IDs [6], in Usenet Message-IDs are required for every post [4]. The *References* header appears when this message is a reply to another post. The contents of this header are one or more Message-IDs of posts in the discussion being replied to. In the experiments performed below, merely the existence of a References header in a Usenet post was enough to consider the message a reply in a discussion thread. The contents of the header were not examined nor were Message-IDs and References headers cross-referenced in any way.

2.1 Custom Newsgroup List

The newsgroups for the hand-chosen newsgroup experiment were taken from high traffic Usenet groups according to Google Groups. Although Google doesn't have a specific ranking of groups, they do categorize groups as "high traffic", "medium traffic", and "low traffic". Also, Google has a very large archive of Usenet, so it is fairly safe to assume that they have enough data to make an accurate determination of the volume of postings per group. The 38 groups chosen were ones that are not used for posting binary files (like those in the *alt.binaries.** tree), contain primarily English postings, and repre-

	Spam	Ham
Non-Replies	15323 (16.6%)	76996
Replies	5299 (1.2%)	420956

Table 1: DSPAM classification of custom newsgroup list

sent a variety of Usenet users and writing styles. For example, messages in *linux.kernel* are likely to contain more technical language than those in *alt.gossip.celebrities*. A full list of groups used for these experiments is provided in Appendix A.

From the list of 38 Usenet groups, all of the available messages were harvested from a local Usenet server. A total of 518,576 messages were harvested, spanning approximately two months.[1] The headers of the posts were discarded, aside from using the References header to determine if the message was a reply as discussed previously. The bodies of non-replies were kept in their original forms, but the bodies of replies were modified to remove quoted text. This was done using a simple pattern matching of lines beginning with common quoting characters like "> " along with a line preceding the quote in a form similar to "On (date) Alice wrote:". The pattern matching was by no means perfect, but it caught a majority of cases. Quoted text was removed because in the case of replies, we are only interested in the new text written by the author. Moreover, we assume that the quoted text in Usenet postings will already be classified separately, since in most cases the original message will be harvested along with any replies.

The results of this experiment are summarized in Table 1. There is a notable difference in the amount of spam in the replies versus non-replies. One of the reasons that the number of replies classified as spam is non-zero is that DSPAM itself is not perfectly accurate: DSPAM has false positives. Likewise, there are also false negatives. However, the greater than ten times increase in the proportion of spam in non-reply postings validates the hypothesis that harvesting only replies provides for a cleaner source of ham text than harvesting all Usenet postings. Another reason for the misclassifications is that some postings, particularly replies, can be very short. Someone may post only one sentence or a couple words (such as "me too"), which does not provide enough data for DSPAM to make an accurate decision. In many cases such messages will be classified as undecided, but if those few words contain some spam-like tokens like "V1@gra" (perhaps as a meme in an otherwise legitimate conversation) then DSPAM will generate a false positive. However, a similar issue exists for false negatives, in that some spam only contains a few words and a URL, none of which may be seen as spammy to DSPAM. That said, if we assume that the goal is to gather ham from Usenet in an automated way, then some false positives and false negatives are to be expected.

In order to verify that the results from running DSPAM were reasonable, we manually classified 2000 messages from the original data set; these were messages that had also been automatically classified by DSPAM. All of the messages were divided

[1]This calculation was based upon message dates, and therefore we assume that clients that generated the date headers were accurate and that the Usenet server expires messages in chronological order.

		DSPAM	
		Spam	Ham
Manual	Spam	33	5
	Ham	467	495

Table 2: Manual classification of replies

		DSPAM	
		Spam	Ham
Manual	Spam	419	155
	Ham	81	345

Table 3: Manual classification of non-replies

into four sets according to whether the message was a reply or non-reply and whether DSPAM classified the message as ham or spam. Within each of these sets, 500 messages were chosen at random using Python's built-in pseudo-random library. This not only chooses a pseudo-random sample of the messages, but also randomizes the order of the sub-list. Although the sample is not a true random sample, it is close enough for this experiment. Each message was then viewed by a human who assigned a verdict of spam or ham to it. A relatively conservative definition of spam was used in this process. Text that didn't flow like usual prose and contained many similar URLs was considered spam. However, messages that appeared to be written by a person but were off-topic to the newsgroup were considered ham. Some posts were a duplication of other published work, such as news articles from web sites, which were also labeled as ham. Other posts contained little more than a URL, so the web site was visited to see if it pointed to a seemingly legitimate web site. Finally, if none of the above factors provided any clue as to whether the message was spam or ham, the message was labeled as ham. This last case occurred in less than 2% of the messages classified in each of the reply and non-reply sets.

The results of the manual classification of each of the sets are provided in Tables 2 and 3. For the reply messages, DSPAM's determination of the amount of spam was considerably higher than the manual classification. While only 1% of the messages that DSPAM classified as ham were in fact spam, 93.4% of messages that it classified as spam were in fact ham.[2] This indicates that the 1.2% spam result shown in Table 1 is probably an upper bound, and could be considerably lower. For non-replies, the results of the manual classification were mixed. DSPAM's classification of non-replies was correct 76.4% of the time, with most of the errors being spam that DSPAM missed, incorrectly classifying the messages as ham. This result could indicate that the amount of spam in non-replies from Table 1 is underestimated. Overall, the manual verification

[2]Interestingly, that 1% (or five messages) of the replies incorrectly identified as ham were, in fact, spam messages that appeared to forge the References header. The Messsage-IDs referenced in the so-called replies could not be found anywhere in the other messages gathered, and the bodies did not contain any text that could be thought to be a quote of a previous message.

	Spam	Ham
Non-Replies	27282 (16.8%)	134909
Replies	55421 (5.8%)	895618

Table 4: DSPAM classification of top 100 newsgroup list

suggests that while DSPAM was not completely accurate, most of the mistakes only strengthen the argument that reply messages contain significantly less spam than non-replies.

2.2 Top 100 Newsgroup List

The second experiment followed exactly the same procedure as the first, except the list of Usenet groups used were taken from NewsAdmin's top 100 text newsgroups list [5], which is a site that gathers statistics on Usenet groups. Some of the groups on the full top 100 list were excluded because they were test groups or spam traps (like *alt.alt.spamtrap*), leaving a final list of 77 groups. This list is given in Appendix B. In total this experiment contained over 1.1 million messages, spanning just over five weeks. The results of this experiment (according to DSPAM) are consistent with the previous one, and are summarized in Table 4.

3 Limitations

In response to the presented method of only harvesting replies as a source of ham, spammers can change their behaviour and start to disguise their spam as replies as well. However, this poses a few problems to the spammer. Many news readers (including Google Groups) group messages together into threads and hide all but the message that started the thread until the user clicks on the thread to view more. Thus, spam messages disguised as replies to existing messages won't be as visible as lone messages, decreasing the probability that a human will see the spammer's message and thereby decreasing the effectiveness of the spam campaign. Also, in order to fool Usenet clients into including spam in an existing thread, the spam message must include a header that references the Message ID of an existing thread. If the spammer randomly generates Message IDs, then Usenet clients will treat the messages as "new" threads. Thus, the spammer must first download recent Usenet postings to harvest the Message IDs, consuming extra bandwidth and computational resources. For both of these reasons, it is unlikely that a spammer would disguise spam as reply messages.

4 Conclusion

Based on the experiments performed, we have shown that Usenet posts can be a good source of ham for evaluating spam filters. While Usenet posts have been used before,

our work shows that extending the idea to only harvest *replies* to other postings provides a good source of ham. Although this source is not perfect, harvesting Usenet replies is a good starting point and can be used for a ham source when a small amount of noise is acceptable.

Acknowledgments

The authors' research was funded in part by the Natural Sciences and Engineering Research Council of Canada; the first author was also supported by the Informatics Circle of Research Excellence.

References

[1] G. V. Cormack and T. R. Lynam. TREC 2005 Spam Public Track Corpora. `http://plg.uwaterloo.ca/~gvcormac/treccorpus/`, 2005. Accessed April 17, 2008.

[2] F. D. Garcia, J. Hoepman, and J. van Nieuwenhuizen. Spam Filter Analysis. In *19th IFIP International Information Security Conference*, pages 395–410, 2004.

[3] Google Inc. Google Groups. `http://groups.google.com/`. Accessed September 12, 2008.

[4] M. Horton and R. Adams. Standard for Interchange of USENET Messages (RFC 1036). `http://www.ietf.org/rfc/rfc1036.txt`, Dec. 1987.

[5] NewsGuy Inc. NewsAdmin / Usenet Statistics / Top 100 text newsgroups by postings. `http://www.newsadmin.com/top100tmsgs.asp`, 2008. Accessed August 6, 2008.

[6] P. Resnick, ed. Internet Message Format (RFC 2822). `http://www.ietf.org/rfc/rfc2822.txt`, Apr. 2001.

[7] J. A. Zdziarski. DSPAM. `http://dspam.nuclearelephant.com/`. Accessed February 19, 2009.

A Custom Newsgroup List

24hoursupport.helpdesk
alt.fan.harry-potter
alt.gossip.celebrities
alt.guitar
alt.home.repair
alt.religion.scientology
alt.support.diabetes
alt.usage.english
comp.lang.c
comp.lang.java.programmer
comp.lang.python
comp.os.linux.advocacy
comp.os.linux.misc
comp.sys.mac.system
linux.kernel
microsoft.public.dotnet.languages.csharp
misc.invest.stocks
rec.audio.pro
rec.autos.sport.nascar
rec.crafts.metalworking
rec.food.cooking
rec.gambling.poker
rec.games.pinball
rec.games.video.arcade.collecting
rec.guns
rec.motorcycles
rec.music.beatles
rec.photo.digital
rec.sport.football.college
rec.sport.golf
rec.sport.pro-wrestling
rec.sport.soccer
rec.travel.europe
rec.woodworking
sci.electronics.design
sci.electronics.repair
sci.math
soc.retirement

B Top 100 Newsgroup List

alt.atheism
alt.comp.freeware
alt.fan.cecil-adams
alt.fan.rush-limbaugh
alt.fifty-plus.friends
alt.humor.puns
alt.marketplace.online.ebay
alt.penthouse.sex.phone
alt.politics
alt.religion.scientology
alt.slack
alt.sports.baseball.bos-redsox
alt.suicide.holiday
alt.support.depression
alt.support.dissociation
alt.tv.american-idol
alt.usage.english
aol.neighborhood.ny.new-york
aol.neighborhood.pa.philadelphia
aol.neighborhood.pa.pittsburgh
aus.politics
comp.lang.labview
comp.os.linux.advocacy
de.talk.tagesgeschehen
fa.linux.kernel
free.alt.freedom.japan.loli
free.uk.tv.bigbrother
fr.misc.engueulades
fr.soc.politique
it.arti.trash
it.comp.console
it.discussioni.animali.gatti
it.discussioni.auto
it.hobby.fai-da-te
it.hobby.motociclismo
it.istruzione.scuola
it.politica
it.politica.internazionale
it.sport.calcio.milan
it.sport.calcio.napoli
linux.kernel
macromedia.dreamweaver
microsoft.public.access
microsoft.public.excel.misc
microsoft.public.excel.programming
microsoft.public.outlook.general
microsoft.public.windows.vista.general
nl.politiek
nz.general
or.politics
pl.misc.samochody
pl.pregierz
pl.rec.fantastyka.sf-f
pl.soc.polityka
pl.soc.prawo
rec.arts.sf.written
rec.boats
rec.food.cooking
rec.gambling.poker
rec.games.pinball
rec.music.artists.springsteen
rec.pets.cats.anecdotes
rec.sport.football.college
rec.sport.pro-wrestling
rec.sport.tennis
sci.physics
soc.culture.israel
soc.retirement
talk.origins
talk.politics.misc
tw.bbs.forsale.house
uk.legal
uk.media.tv.misc
uk.people.silversurfers
uk.politics.misc
uk.rec.motorcycles