# Towards a Dataset for the Discrimination of Warranted and Unwarranted Spam


Colorado State University

Rays Cyber Research Lab

Eric Burton Samuel Martin

Advisor: Indrakshi Ray  Co-Advisor: Hossein Shirazi

SDSU San Diego State University

Center for Cybersecurity Analytics and Automation


## Goals:

- Increase the effectiveness of current spam filters that rely on user-feedback
  - Decrease malicious spam from bypassing filters
  - Decrease spam filter bias against commercial correspondence
- Produce a model that can discriminate between warranted and unwarranted spam

## Motivation:

Affiliate marketing and survey-based spam has been bypassing modern spam filters

*"Affiliate marketing and survey-based spam have been consistently bypassing our corporate spam filters"*

---

**What incentive do spammers have to send these?**

- To collect affiliate commissions, obtain and sell user-data, and can deploy malware

**Brands associated with these unsolicited spam emails**

- Risk tarnished reputations and begin to be perceived as 'spammy'

## Background on Spam:

"unsolicited", "irrelevant", "inappropriate", "unwanted", "commercial messages", "advertising material"

**There are numerous definitions of spam, but here is Merriam-Webster's definition :**

**spam** 1 of 3 **noun**

'spam

: unsolicited usually commercial messages (such as emails, text messages, or Internet postings) sent to a large number of recipients or posted in a large number of places

**Most perceived "spam" is solicited and benign:**

- A significant portion of emails perceived as spam are messages that recipients have, in some manner, consented (explicitly or implicitly) to receive
  - Explicit consent (signing up to a newsletter)
  - Implicit consent (terms and conditions, making online purchases, etc.)
- These emails, although unwanted, often abide by the country of origins spam laws

**Unsolicited spam is a risk to individuals and businesses:**

- Unsolicited spam can be used to carry out phishing attacks, distribute malware, perpetuate scams and fraud, collect user-data for identity theft, and to collect affiliate commissions

## Conjecture:

Why spam filters that rely on user-feedback to train their models may be the cause of this issue

**Reliance on User Feedback**

- Users tend to flag solicited emails as spam rather than unsubscribe
- Risk of introducing noise to spam filter model due to incorrect spam labeling
- Potential to over-weight patterns of authentic commercial correspondence
- Here is how Google's spam filters claim to work:

Simply put, to protect users at scale, we rely on machine learning powered by user feedback to catch spam and help us identify patterns in large data sets—making it easier to adapt quickly to ever-changing spam tactics. Gmail employs a number of AI-driven filters that determine what gets marked as spam. These filters look at a variety of signals, including characteristics of the IP address, domains/subdomains, whether bulk senders are authenticated, and user input. User feedback, such as when a user marks a certain email as spam or signals they want a sender's emails in their inbox, is key to this filtering process, and our filters learn from user actions. [1]



**Bias in Training Data**

- Model effectiveness tied to data quality
- Bias against legitimate emails due to user misreporting
- Leads to higher false positive rates for commercial mail

**Impact on Spam Detection Effectiveness**

- Risk of creating blind spots in detection
- Over-tuning to user-reported spam can miss subtle spam cues
- Real malicious unsolicited spam may slip through due to learned biases

**Nuanced Approach for Effective Spam Detection**

- Introduce two new classification labels for spam 'warranted' and 'unwarranted'
- Generate a model to distinguish between these two classifications
- Classify all user-flagged spam as 'warranted' or 'unwarranted'
  - If warranted, do not use to update spam-filter model, add rule to filter
  - If unwarranted, use in future spam-filter model updates

## 'Warranted' Spam and 'Unwarranted' Spam:

Introducing nuanced classification for spam

To aid in the discrimination of unsolicited and solicited spam, we introduce two labels:

**'Warranted' Spam:**

*Legitimate communications that recipients have consensually opted into, knowingly or not, that originate from a credible source, and that provide clear and safe opt-out methods.*

**'Unwarranted' Spam:**

*Unsolicited and often malicious messages sent without the recipient's consent, where attempts to unsubscribe may be futile or may even exacerbate the problem.*

There exists a public dataset that matches the unwarranted spam definition, the Spam Archive by Bruce Guenter [2] . But there is no warranted spam dataset. So, we made one.

## Limitations of Available Spam Datasets:

**Currently Available Spam Datasets:**

| Dataset | Age of Data |
|---|---|
| Ling-Spam | 2000 |
| SpamAssassin | 2000-2006 |
| Enron-Spam | 2006 |
| TREC07 | 2007 |
| CSDMC | 2010 |
| Bruce Guenter | 1997-2023 |

- **Outdated, Outdated, Outdated!**
  - Aside from the Bruce Guenter dataset, every dataset is unacceptably outdated
  - Models trained on these datasets show deteriorated performance when tested on current data. [3]
- **May contain Amalgamation of unwarranted and warranted spam**
  - Can therefore contain user-flagged warranted spam
- **The exception is the Bruce Guenter dataset [2]**
  - Consists solely of unwarranted spam sent to a bait address

## Introducing the Warranted Spam Dataset:

**We have created a large dataset consisting of warranted spam**

- Up-to-date, Modern, Ever-growing, public facing warranted spam archive [4]

**The dataset is comprised of two main email account repositories:**

1) PRIMARY@gmail.com
   - The primary dataset (meticulously managed and tracked)
   - Logs all websites the address is registered with into our website repository
   - Employs the '+' feature in Gmail to trace the original source of each email
     - For example, PRIMARY+nike@gmail.com is used to register for Nike.com
2) AD-HOC@gmail.com
   - A dataset designed for convenient, ad-hoc signups outside of working hours
   - Does not maintain a record of each sign-up
   - Does not utilize the '+' feature

**Website Repository:**

- Spreadsheet containing all websites the primary account has registered with
- 28 categories | > 70 websites per category | Unique email address used per site
  - Category examples are finance, news, retail, cannabis, sports, survey, etc.

| AC | AD | AE | AF | AG |
|---|---|---|---|---|
| Fashion | | | | |
| Fashion_Company | Fashion_URL | Fashion_Registration_Email | Done | News_Company |
| Vogue | https://www.vogue.com/ | REDACTED+vogue@gmail.com | ☑ | BBC News |
| Elle | https://www.elle.com/ | REDACTED+elle@gmail.com | ☑ | CNN |
| Harper's Bazaar | https://www.harpersbazaar.com/ | REDACTED+harpersbazaar@gmail.com | ☑ | The New York Times |

**Sign-up Methodology:**

- Our team has manually attempted to register to over 2000 websites so far
- Automation of the sign-up process has been attempted with low success rates

**Data Cleansing and Management:**

- We have scripts to export, organize, scrub, and zip all email files
- Key metrics are extracted from each raw email for future use
- All mail sent to the Gmail Spam folder is redirected to a 'Labelled Spam' folder

| | Primary | Ad-Hoc | Forwarded |
|---|---|---|---|
| Provider | Gmail | Gmail | Outlook |
| Instantiation | 3-May-23 | 31-Mar-23 | 18-May-23 |
| GB | 14.12 | 4.93 | 15 (Max) |
| Total Emails | 164.8K | 71.2K | 54.4K |
| Spam | 1.4K (2.0%) | 1.4K (2.90%) | 1.4K (2.60%) |

**Table 1: Summary of statistics from creation date until 12-Nov-2023 for each email Account used in dataset collection.**

## Next Steps:

Utilize the dataset and leverage large language modes and NLP

**Distinguish between unwarranted and warranted spam:**

- Utilize the Bruce Guenter Spam dataset as unwarranted spam and our dataset as warranted spam to train a model to perform classification

**Content categorization:**

- Utilize LLMs and NLP to categorize warranted spam into categories, such as News, Finance, Sports, Fashion, etc.

**Sentiment Analysis:**

- Compare warranted spam vs unwarranted spam to see similarities and differences

## Use Case Scenario:

Harden spam filters against unwarranted spam

When a user flags an email as spam, run it through our classification filter, then:

- **If the flagged email is classified as 'warranted spam'**
  - Add a rule to prevent this mail from entering the user's inbox
  - Do not use this mail to update the spam-filter machine learning model
- **If the flagged email is classified as 'unwarranted spam'**
  - Add a rule to prevent this mail from entering the user's inbox
  - Use this mail to update the spam-filter machine learning model

## References:


[1] https://workspace.google.com/blog/identity-and-security/an-overview-of-gmails-spam-filters

[2] https://untroubled.org/spam/

[3] Jáñez-Martino, F., Alaiz-Rodríguez, R., González-Castro, V. et al. A review of spam email detection: analysis of spammer strategies and the dataset shift problem. Artif Intell Rev 56, 1145–1173 (2023). [4]

[4] https://www.cs.colostate.edu/~ebmartin/warrantedSpamDataSet/