Towards a Comprehensive Dataset for Discrimination between Warranted and Unwarranted Emails

Eric Burton Samuel Martin
eric.burton.martin@colostate.edu
Department of Computer Science
Fort Collins, CO, USA

Hossein Shirazi
hshirazi@sdsu.edu
Fowler College of Business
San Diego, CA, USA

Indrakshi Ray
indrakshi.ray@colostate.edu
Department of Computer Science
Fort Collins, CO, USA

Abstract

In this research, the prevailing issue we address is the overgeneralized perspective of spam/ham (non-spam) classification. Despite the intricacies of spam classification, reliance on user feedback may inadvertently skew filters to misclassify legitimate and malicious email, as users are prone to flag innocuous commercial mail as spam rather than unsubscribing. Current spam datasets have a propensity to include such user-flagged spam which can lead to further misclassification, leading to filters biased against warranted commercial correspondence. Motivated to address this concern, we introduce two new classification categories that delve deeper into the nuances of spam. 'Warranted spam', refers to consensual communications, from a credible source with transparent and safe opt-out mechanisms, and 'unwarranted spam' describes unsolicited messages, often of a malicious nature. Utilizing these classifications, we propose an innovative and dynamic 'warranted spam' dataset that seeks to pave the way for researchers to develop more sophisticated spam filtering techniques. Furthermore, our study delves into pioneering machine learning and natural language processing approaches, harnessing our dataset's potential. The overarching aspiration of our work is to augment online safety, preserve brand integrity, and optimize both the user experience and the efficacy of email marketing campaigns.

CCS Concepts: • **Information systems** → **Document filtering**; **Information extraction**; • **Security and privacy** → *Information flow control*; **Usability in security and privacy**; **Economics of security and privacy**; **Social aspects of security and privacy**; • **Computing methodologies** → *Machine learning*.

Keywords: Warranted Spam, Unwarranted Spam, Spam Detection, Dataset, Spam, Email Filtering, Spam Classification, Marketing Emails, Machine Learning, Digital Communication

ACM Reference Format:
Eric Burton Samuel Martin, Hossein Shirazi, and Indrakshi Ray. 2023. Towards a Comprehensive Dataset for Discrimination between Warranted and Unwarranted Emails. In *Proceedings of Nov 26–30, 2023 (ACM CCS '23').* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

ACM CCS '23', Copenhagen, DK,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Currently, spam email is broadly defined as any unsolicited message, typically of commercial origin. However, is this definition, centered around the notion of 'unsolicited', inadvertently limiting the effectiveness of our spam filtration methodologies? The term 'unsolicited' implies messages that are "not asked for or requested' In fact, a significant portion of emails perceived as spam are messages that recipients have, in some manner, consented to receive. This consent may be explicit, as in the case of newsletter subscriptions, or implicit, as when individuals agree to terms and conditions, make online purchases, or access free online services. Oftentimes in these cases, users inadvertently agree to receive future communications. These emails, although potentially unwanted, are warranted in the sense that they are sent with some level of prior consent. This paper proposes a distinction be made between 'warranted spam' and 'unwarranted spam'. We define 'warranted spam' as legitimate communications that recipients have opted into, that originate from a credible source, and that provide clear and safe options for recipients to opt-out. Whereas 'unwarranted spam' is defined as unsolicited and often malicious messages sent without the recipient's consent, where attempts to unsubscribe may be futile or may even exacerbate the problem. Current spam datasets available to researchers, such as Ling-Spam [2], SpamAssassin [8], Enron-Spam [7], TREC07 [3], CSDMC [1], and others, contain data that is over a decade old which, according to Jáñez-Martino et al. [5], has significant implications for researchers. They show that models trained on these outdated datasets often have deteriorated performance when tested on current data. Along with being outdated, most, if not all, of these datasets consist of an amalgamation of warranted and unwarranted spam. We propose that this broad categorization of spam hampers the development of accurate and effective filtering solutions, as it forces the characteristics of warranted and unwarranted spam to be clumped together.

This is a significant issue, as unwarranted spam is not just a nuisance but a vector for phishing attacks, malware distribution, and other cyber threats. Moreover, unwarranted affiliate marketing spamming mechanisms are increasingly leveraged for financial advantage as spammers are incentivized to distribute unwarranted spam to accrue affiliate marketing commissions. This can lead to unintended consequences for the companies whose reputations and brand identities are being affiliated with unwarranted spam. Recognizing this, we are currently developing a modern dataset consisting of warranted spam. This dataset is meticulously curated through manually signing up for email correspondence as a normal user would. Our dataset is designed to complement existing unwarranted spam datasets, such as the one maintained by Bruce Guenter [4], allowing for a more complete and comprehensive understanding of spam. For individual users, this nuanced approach to spam detection promises reduced exposure to malicious content. For businesses, especially those that rely on email marketing, our dataset offers potential for improved deliverability of communications, increased customer engagement, and compliance with various regulations, such as the CAN-SPAM Act and GDPR. Importantly, it also helps to preserve brand integrity, as companies' legitimate communications are less likely to be mistakenly classified as spam, which can tarnish their reputation and erode customer trust.

2 Approach for Dataset Generation

The dataset generation process aims to collect warranted spam emails.

2.1 Warranted Spam Email Accounts

Three separate email accounts (actual addresses withheld for confidentiality) were created for collecting warranted spam emails.

1. **TRACKED@gmail.com** - The primary email account. It employs the '+' feature in Gmail to trace the original source of each email. This account is rigorously managed and actions are meticulously documented.
2. **UNTRACKED@gmail.com** - A supplementary email account. Unlike the tracked account, this dataset does not maintain a record of each sign-up. It is designed for convenient, ad-hoc sign-ups encountered during researchers' daily internet use outside of working hours.
3. **FORWARD@outlook.com** - This account receives emails forwarded by the tracked Gmail account. This incorporates email headers generated by Microsoft Outlook into the dataset.

It should be noted that all emails received that are categorized as spam by the email provider are forwarded to a dedicated "Labelled Spam" folder for easy identification and segregation for future analysis.

2.2 Website Repository

We established a diverse website repository [6], comprising over 28 categories including retail, finance, health, sports, food, beauty, fashion, news, crypto, and more, each consisting of over 70 websites generated mostly utilizing ChatGPT4 May 3 Version. Table 1 lists these topics. The repository tracks the website URL, the unique email address provided to the site (utilizing the '+website@gmail.com' feature, and whether registration was successful with accompanying descriptive comments.

Retail, Travel, Finance, Health, Sports, Food, Beauty, Fashion, News, Crypto, Wedding/Party, Parenting, Home Improvement, Cannabis, Automotive, Real Estate, Software as a Service, Coupons, Survey/Prize, Arts, Education, Entertainment, Science, Music, Religion, Self-Help, Freelancer/Get Rich, Job Search

Table 1. List of categories in the website repository.

2.3 Signup Methodology

The websites outlined in the repository 2.2 were systematically chosen to maximize affiliate communications, marketing emails, and newsletters. Each site in the repository is manually visited and surveyed and a unique email was supplied if possible. This usually included newsletter sign-ups, pop-ups, or account creation procedures that included an opt-in email system. Although enticing, automation of the sign-up process has proven infeasible due to the diverse and complex registration requirements across different websites, including CAPTCHA challenges, phone number verifications, and unique form structures designed to deter automated interactions.

2.4 Data Cleansing and Management

We used several Python scripts for data management and cleansing. These scripts perform tasks such as extracting individual emails from large .mbox files, organizing emails chronologically, scrubbing sensitive recipient data to maintain anonymity, and updating the website [6] to reflect the most recent additions.

2.5 Current Dataset Statistics

We noticed a substantial difference in storage consumption between the tracked Gmail account and its corresponding Outlook account, which receives emails forwarded from the tracked account. Even with a comparable number of emails, the storage used by Outlook significantly exceeds that of Gmail and has reached the 15 GB maximum storage in a matter of months. Our aim is to investigate the underlying reasons for this discrepancy. Table 2 highlights some base statistics regarding our datasets at the time of writing.

	Tracked	Untraced	Forwarded
Provider	Gmail	Gmail	Outlook
Instantiation	3-May-23	31-Mar-23	18-May-23
GB	6.12	1.93	15 (Max)
Total Emails	60.8K	23.2K	54.4K
Spam	1.2K (2.0%)	0.6K (2.90%)	1.4K (2.60%)

Table 2. Summary of statistics from creation date until 20-Aug-2023 for each email Account used in dataset collection.

3 Next Steps

Our proposed system seeks to introduce a more comprehensive labeling mechanism that will:

- Distinguish between unwarranted and warranted spam.
- Identify potential misclassifications in public datasets.
- Perform content categorization and sentiment analysis.

Our approach is divided into two main components:

Leveraging Large Language Models (LLM). By leveraging LLMs, we aim to achieve a high degree of accuracy in distinguishing between warranted and unwarranted spam by delving into the textual content of emails. This involves a multifaceted approach: firstly, by discerning semantics, context, and recurring patterns within the email body, we strive to enhance the efficiency of our spam detection. Beyond mere detection, the LLMs facilitate in-depth content categorization, which proves instrumental in determining the intrinsic nature and underlying purpose of the email, thereby allowing a more nuanced labeling mechanism. Lastly, the models are adept at sentiment analysis, extracting the emotional undertones and intentions encapsulated within the email's content. Such analysis proves invaluable, particularly when it comes to identifying potential phishing attempts or scam emails that exploit recipients' emotions.

Analysis of Metadata and Feature Engineering. While LLMs help in deciphering and classifying textual content, it's crucial to acknowledge the significance of the email's non-textual components. Spammers often resort to tactics like crafting bespoke headers or manipulating other metadata attributes, all in a bid to elude conventional filters. In light of this, our strategy places considerable weight on a rigorous metadata analysis designed to pinpoint anomalies or characteristic patterns synonymous with spam emails. Central to this strategy is the aspect of feature engineering. Through meticulous selection and transformation of raw data extracted from email headers and associated metadata into a cohesive set of features, we anticipate enhancing the precision with which our system identifies and classifies spam. Furthermore, these meticulously engineered features, when synergized with traditional classifiers or harmoniously integrated with LLMs, promise to substantially fortify our spam detection prowess.

This holistic perspective is expected to substantially elevate the precision and accuracy of spam detection.

4 Dataset Use-Cases and Implications

4.1 Security-Related Uses

- **Phishing/Spam Detection:** By pairing this dataset with unwarranted datasets like Guenter's [4], researchers can train models to discern legitimate marketing emails from phishing/spam attempts.
- **Safe Unsubscription:** Examining the opt-out mechanisms in warranted spam can help create safer unsubscription tools, shielding users from threats when unsubscribing.

4.2 Business and Marketing Uses

- **Identifying Spam Triggers:** By evaluating the warranted spam flagged as spam by Gmail, businesses can pinpoint elements triggering spam filters, refining their communication strategies.
- **Temporal Analysis:** The dataset's chronological structure allows tracking of email marketing trends and volume over time.

The implications of distinguishing warranted from unwarranted spam are given below.

4.3 Improved User Experience

- **Enhanced Security:** Better filtering of unwarranted spam reduces user exposure to malicious content, bolstering online safety.

4.4 Enhanced Email Marketing Effectiveness for Businesses

- **Improved Deliverability:** Ensuring emails from businesses are classified as warranted increases the likelihood of reaching intended recipients, avoiding spam folders.
- **Brand Integrity:** For legitimate companies, having their brand unknowingly associated with unwarranted spam can tarnish their reputation.

References

[1] 2010. CSDMC2010 SPAM Corpus. http://csmining.org/index.php/spam-email-datasets-.html.

[2] I. Androutsopoulos et al. 2000. Ling-Spam Dataset. https://www.kaggle.com/datasets/mandygu/lingspam-dataset.

[3] G. V. Cormack. 2007. TREC07 Spam Corpus. In *TREC 2007 Spam Track*.

[4] B. Guenter. 1997-2023. Spam Archive. http://untroubled.org/spam/. Accessed: June 2021.

[5] F. Jáñez-Martino, R. Alaiz-Rodríguez, V. González-Castro, et al. 2023. A review of spam email detection: analysis of spammer strategies and the dataset shift problem. *Artif. Intell. Rev.* 56 (2023), 1145–1173. https://doi.org/10.1007/s10462-022-10195-4

[6] E Martin. 2023. Warranted Spam Archive. https://www.cs.colostate.edu/~ebmartin/warrantedSpamDataSet/. Hosted by Colorado State University.

[7] V. Metsis et al. 2006. Enron-Spam Dataset.

[8] Apache SpamAssassin Project. 2005. SpamAssassin. https://spamassassin.apache.org/old/publiccorpus/.