Decoding Email Strategies: Classification and Categorization of Warranted and Unwarranted Spam

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

ABSTRACT

PROJECT PROPOSAL ABSTRACT

Building upon my prior research, which introduced a nuanced approach to spam classification and supplied a brand-new 'warranted spam' dataset, I will be working towards addressing the future work section by exploration the realm of email content analysis using Large Language Models (LLMs) and Natural Language Processing (NLP) techniques. My previous work highlighted the challenges posed by user-driven spam flagging and the subsequent biases introduced in spam filters. With the foundation of the 'warranted spam' dataset, combined with the Bruce Guenter "unwarranted spam" dataset, I am poised to delve deeper into the textual intricacies of these emails.

My primary objective is to harness the capabilities of LLMs and/or NLP techniques to differentiate between warranted and unwarranted spam by analyzing the textual content of emails. LLMs, with their ability to discern semantics, context, and recurring patterns, present a promising avenue for effective classification of warranted and unwarranted spam. Since the warranted spam dataset is composed of commercial correspondence that falls in 28 categories, I aim to employ LLMs and/or NLP techniques for in-depth content categorization of emails, providing insights into the intent and nature of emails. This categorization will aid filters and users in understanding the underlying motivations behind emails, especially when comparing the sentiment of warranted commercial communications against potential phishing or scam emails.

By juxtaposing the sentiments of these two distinct categories, I hope to shed light on the emotional tactics employed by both legitimate marketers and malicious actors. The ultimate goal of this research is to refine spam filtering techniques, ensuring that genuine communications are accurately identified and easily categorized, while also safeguarding users from potential threats. This endeavor not only seeks to enhance online safety but also aims to provide businesses with actionable insights to refine their email communication strategies.

THE REST OF THIS DOCUMENT IS MY PREVIOUS WORK

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*; *Natural language processing*.

KEYWORDS

Warranted Spam, Unwarranted Spam, Spam Detection, Dataset, Spam, Email Filtering, Spam Classification, Machine Learning, Natural Language Processing

ACM Reference Format:
Eric Burton Samuel Martin. 2023. Decoding Email Strategies: Classification and Categorization of Warranted and Unwarranted Spam. In *Natural Language Processing, August 21 – December 13, 2023, Fort Collins, CO, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

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 consensually opted into, knowingly or not, 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 [7], Enron-Spam [6], 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.

CS542 '23, August 21 – December 13, 2023, Fort Collins, CO, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

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 by 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 DATASET GENERATION METHODOLOGY

The dataset generation process aims to collect a large quantity of warranted spam emails. We are providing three datasets to the public in our Warranted Spam Archive[1].

2.1 Warranted Spam Email Datasets

To host our warranted spam, we have created three accounts.

1. **PRIMARY@gmail.com** - The primary dataset. This account meticulously registers to websites found in the website repository[1] and employs the '+' feature in Gmail to trace the original source of each email. For example, PRIMARY+nike@gmail.com would be used to register for Nike.com.
2. **AD-HOC@gmail.com** - A supplementary dataset that, unlike the primary account, does not maintain a record of each sign-up. This dataset is designed for convenient, ad-hoc sign-ups encountered during researchers' daily lives outside of working hours.
3. **FORWARD@outlook.com** - This account has the single purpose of receiving emails forwarded by the primary Gmail account. This incorporates email headers generated by Microsoft Outlook into the dataset.

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. This allows researchers to investigate features that can cause commercial warranted spam to be classified as spam.

2.2 Website Repository

For the primary email account, we created a diverse website repository[1] spanning 28 categories that were systematically chosen to maximize affiliate communications, marketing emails, and newsletters. Each category contains more than 70 websites that were generated utilizing ChatGPT4 and Google and vetted by the team. A few examples of the categories within the repository are Retail, Travel, Finance, Health, Sports, Food, Beauty, Fashion, News, Crypto, and Job Search. The repository contains the website name, URL, unique email address provided to the site (using the '+' feature), whether registration was successful, and comments if applicable.

2.3 Sign-up Methodology

For the primary account, each site in the website repository[1] was manually visited and surveyed for newsletter sign-ups, pop-ups, or account creation procedures that included an opt-in email system. Once a place for an email address was found, a unique email address employing the '+' feature was supplied and if possible, the highest email frequency was chosen. For the ad-hoc account, researchers provide the email address when they encounter a sign-up form in their daily lives. The ad-hoc account does not use the website repository as a reference. Therefore, the primary and ad-hoc accounts may have some overlapping sign-ups.

2.4 Sign-up Automation

Automation of the sign-up process has proven difficult due to the diverse and complex registration requirements across different websites, including CAPTCHA challenges, phone number verification, and unique form structures designed to deter automated interactions. We are currently working on an automatic sign-up bot for websites that have simple email input fields to help dramatically increase our dataset size.

2.5 Data Cleansing and Management

We developed 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, extracting key metrics from collected emails, including sender domain, unsubscribe link count, presence of tracking pixels, and authentication results, among others, and updating the warranted spam archive website [1].

We noticed a substantial difference in storage consumption between the primary Gmail account and its corresponding Outlook account, which receives emails forwarded from the primary account. Even with a comparable number of emails, the storage used by Outlook significantly exceeds that of Gmail as seen in Table 1.

	Primary	**Ad-Hoc**	**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 1: Summary of statistics from creation date until 20-Aug-2023 for each email Account used in dataset collection.

[1] EbMartin. 2023. Warranted Spam Archive. https://www.cs.colostate.edu/~ebmartin/warrantedSpamDataSet/

3 NEXT STEPS FOR CLASSIFICATION

Utilizing our dataset, we are working towards the following:

- Distinguish between unwarranted and warranted spam.
- Identify potential misclassifications in public datasets.
- Content categorization and sentiment analysis.
- Automation of email registration.
- Unsubscription link and image link study.

Our approach towards the classification of warranted and unwarranted spam is divided into three main components:

3.1 Leveraging Large Language Models (LLMs)

Through the use of LLMs, we aim to achieve differentiation between warranted and unwarranted spam by analyzing the textual content of emails. LLMs allow us to discern semantics, context, and recurring patterns, facilitating effective classification. Moreover, they enable in-depth content categorization, offering insights into the email's intent and nature. Additionally, we intend to employ sentiment analysis to gauge emotional undertones and motivations within the email content, especially comparing the sentiment of commercial warranted spam against potential phishing or scam emails since they both often attempt to exploit recipients' emotions to generate clicks.

3.2 Analysis of Metadata and Feature Engineering

Beyond textual content, the non-textual components of emails play an important role in spam detection. Spammers often resort to tactics like crafting bespoke headers and manipulating metadata attributes, in an attempt to elude conventional filters. In light of this, we have created a complex feature extracting script that will provide us with a vast set of features extracted from email headers and associated metadata. By identifying and assessing patterns unique to both warranted and unwarranted spam, we aim to establish a 'fingerprint' for warranted spam, assisting in the detection of anomalies in unwarranted emails. Utilizing this technique of anomaly detection can help fight against the consistent adversarial techniques spammers use to manipulate metadata attributes to bypass filters abd when synergized with traditional classifiers and LLMs, show promise to substantially fortify our spam detection methods.

3.3 Analysis of Warranted Spam Characteristics

A core aspect of our research is identifying characteristics of warranted commercial spam. We are investigating the frequency and authenticity of unsubscription links in these emails. By understanding how often these links appear and verifying their legitimacy, we aim to highlight potential differences between genuine marketers and malicious entities. Additionally, given that unwarranted spam often heavily relies on images, we are assessing the quantity of images, the existence of embedded links within them, and the proportion of accompanying text in warranted commercial spam. The goal is to ascertain if there's a distinct pattern in how legitimate entities use images in their communications compared to unwarranted spammers.

4 DATASET USE-CASES

A few of the potential use-cases for the warranted spam dataset are proposed below.

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 that aim to mimic these legitimate emails.
- **Safer Unsubscription:** Examining the opt-out mechanisms in warranted spam can help create safer unsubscription tools, shielding users from threats when unsubscribing.

4.2 Business Benefits

- **Improved Deliverability:** Ensuring emails from businesses are classified as warranted increases the likelihood of reaching intended recipients, avoiding spam folders.
- **Identifying Spam Triggers:** By evaluating the warranted spam flagged as spam by Gmail, businesses can pinpoint elements triggering spam filters, refining their communication strategies.
- **Brand Integrity:** For legitimate companies, having their brand unknowingly associated with unwarranted spam can tarnish their reputation. Utilizing this dataset can help train models to identify warranted and unwarranted affiliate marketing.

5 THREATS TO VALIDITY

The integrity of our dataset relies on the websites used to amass warranted spam emails. While we sought diversity in our website selection, most were generated via ChatGPT4 (May 3 Version), so there's potential for geographical, cultural, or other biases that might not capture the global spectrum of spam characteristics. Ensuring the confidentiality of the dataset's email accounts remains crucial to prevent unintended external influences, which could adulterate the dataset. Our pursuit of automation also may introduce bias as our bot will have varying constraints, limiting its scope.

ACKNOWLEDGMENTS

This work was supported in part by funding from NSF under Award Numbers DMS 2123761, CNS 1822118, and from AFRL, ARL, Statnett, AMI, NewPush, and Cyber Risk Research and from NIST under Award Number 60NANB23D152.

We extend our gratitude to Evan Anspach, Alex De La Chica, and Tobias Beesley at Colorado State University, Jhin Echon and Adrian Castro Farias at San Diego State University, Emmanuel Gonzalez at California State University, Los Angeles, and the talented high school student Udhirna Krishnamurthy, for their invaluable contributions to the manual sign-up process, which played a pivotal role in the creation of this dataset.

We would also like to acknowledge Colorado State University for the resources and facilities provided by the institution.

Special thanks to the RaysCyberLab for their continuous guidance, time, insight, and expertise to enhance this research.

REFERENCES

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

[2] I. Androutsopoulos, J. Koutsias, K.V. Chandrinos, George Paliouras, and C.D. Spyropoulos. 2000. Ling-Spam 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] V. Metsis. 2006. Enron-Spam Dataset.

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