Towards the Discrimination of Warranted and Unwarranted Spam

Eric Burton Samuel Martin

eric.burton.martin@colostate.edu

Department of Computer Science

Fort Collins, CO, USA

CS 542 – NLP

Presentation 05

SDSU | San Diego State University

Center for Cybersecurity Analytics and Automation

Colorado State University

Rays Cyber Research Lab

Goal:

Increase the effectiveness of current spam filters that rely on user-feedback

Motivation:

"Malicious affiliate marketing and survey-based spam have been consistently bypassing our corporate spam filters"

Background on Spam:

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

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

Conjecture:

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 [1] emails in their inbox, is key to this filtering process, and our filters learn from user actions.

Bias in Training Data

- Model effectiveness tied to data quality
- Bias against legitimate emails due to user misreporting

Impact on Spam Detection Effectiveness

- Over-tuning to user-reported spam can miss subtle spam cues
- Real malicious unsolicited spam may slip through due to learned biases

‘Warranted’ Spam and ‘Unwarranted’ Spam:
Introducing nuanced classification for spam

Nuanced Approach for Effective Spam Detection

- Introduce two new classification labels for spam ‘warranted spam ’ and ‘unwarranted spam’

‘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.

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 Dataset:

The dataset we made

With the power of undergraduate researchers, we have created a large dataset of warranted spam

- A meticulously managed, ever-growing, modern dataset that contains warranted spam

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

Dataset is publicly available:

https://dl.acm.org/doi/10.1145/3576915.3624397

	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.

Unwarranted Spam Dataset:

The dataset of unwarranted spam

Bruce Guenter's Spam Archive

- Posted a bait address to various forums and has been collecting mail from spammers who scrape for addresses

https://untroubled.org/spam/

Classification Model:

1. **Data Preprocessing:** Clean and manipulate the dataset and generate features
   - **Textual features**: Email Body, Email Subject
   - **Non-textual features**: # Links, # Unsubscribe Links, Tracking Pixel, # HTML tags, tag_to_text_ratio, max_nesting_depth, and more

2. **Tokenize:** Utilize BigBird tokenizer to tokenize textual features for use in RNN/LSTM model

3. **RNN/LSTM Layer**: Processes sequential text data, capture temporal dependencies

4. **BigBird Encoder**: Output of RNN/LSTM is fed into the BigBird encoder to enhance understanding of text

5. **Non-Textual Feature Processor**: A simple feedforward neural network dedicated to analyzing non-textual features like unsubscribe link count and email size.

6. **Concatenation Layer**: Merges outputs from BigBird and the non-textual feature processor.

7. **Decision Layer**: A fully connected neural network layer making the final classification based on combined textual and non-textual insights.

8. **Classify!**

Results: