Spamalot II: Spam Bothers Me

Machine Learning – Defining & Identifying Spam

Christopher Choy, Bijan Oviedo, Scott Wang, Steven Wei
The University of Chicago, CMSC 23210 - Usable Security and Privacy, Spring 2018
{cchoy, boviedo, zirenw, srwei} @uchicago.edu

ABSTRACT

I*n this paper, we develop a study for the purpose of improving machine learning classifiers on graymail–email messages that could be considered spam or not spam by different email users. We conduct a pilot survey study on 7 participants by asking them questions about emails from their private Gmail inbox. We iterate on the survey design and develop a method of automating the conduct of the survey on a larger scale for potential use in the future. In our analysis of the pilot survey data, we find that the more a user receives emails from a sender, the higher the probability she would consider another email from this sender as spam. We conclude by discussing our study's limitations as well as its potential future use cases.*

1. INTRODUCTION

The battle between spam creators and spam filter identifiers grows ever more complex. Almost all present-day electronic mail (abbv. email) spam filters utilize some form of machine learning algorithm, with the algorithms themselves differing from platform to platform. However, despite all effort put forth, current email spam filters are not perfect; there exist many user-defined spam emails that are not caught by current machine learning algorithms. To aide the development of spam email filters' machine learning algorithms, our research aims to build a pilot study for a future, large-scale study into identifying novel classifiers for these algorithms by examining how users practically define and identify spam within their own inboxes.

In this study, we conduct a survey to characterize the emails that users receive and find correlations between these characteristics and whether or not users identify them as spam. To that end, we create and conduct an in-person pilot study on 7 participants gathered from word-of-mouth. To ground this study in participants' own lives and interactions, we ask questions about 10 emails within their own personal Gmail inbox.

Our survey consists of three parts. The first part asks non-intrusive demographic questions. The second part asks generic questions regarding participant feelings towards spam and spam filters as well as their usage frequency of Gmail. The third part asks participants questions whose responses were specific to an email from participants' own inbox. In particular, we investigate how much of the email's contents they actually read through, whether they consider the email to belong in their inbox, and how they regard the sender of the email.

After conducting the survey on each participant, we would make improvements to its contents based on feedback given by participants. This included clearing up ambiguously-worded questions, modifying categories, and adding or removing certain questions. In the end, we create an online survey with polished questions ready to be used in a future iteration of this study.

Our study focuses on developing a method through which new spam-identifying machine learning classifiers could be created based upon user input grounded in users' own Gmail inboxes. From the data gathered in our pilot study, we articulated and demonstrated how to analyze this data in ways that could yield relevant and interesting results.

2. RELATED WORKS

In this section we examine the related work in this research space by considering three questions.

How do users perceive spam?

The effects of spam emails on user productivity has been studied extensively in quantitative and qualitative studies in both a personal setting and an organization setting. [1-7] One study found that spam email messages may cost businesses approximately 1200 minutes per employee per year. [2] In another study from 2003, a survey of 204 email users found that although respondents receive a large amount of unsolicited emails they find annoying, most respondents spend very little time dealing with them, being unlikely to open or read, let alone respond to unsolicited emails. [4] A separate study found that approximately most users spend at least 5 minutes per day on spam, with 32% of users spend at least 15 minutes per day on spam. This same study found that while spam email makes up different proportions of people's personal inboxes across different population subgroups, users' perception of spam–that is, how annoying they find it–is relatively constant for every type of user. However, spam seems to be more a problem for people in their personal email accounts than in their work email accounts. [7]

What are the current methods for detecting spam?

Identification of unsolicited email messages is an ongoing and popular field of research. Traditional techniques used for filtering spam include rule based filtering and distributed adaptive blacklists, yet the overwhelming focus of current research is on machine learning techniques such as Naive Bayes, K nearest neighbors, support vectors, ensemble methods, and multi-layered neural networks. [8-17] Many of these machine learning methods have error rates on untrained spam of less than 5%, with state of the art systems having error rates of less than 2.5%. [24]

Yet the problem of spam detection and classification continues to increase in difficulty as spammers personalize spam messages to individual users. A study conducted in 2017 by Ezpeleta *et al.* showed that spammers use users' public information on Facebook to personalize spam email messages, specifically to tailor the subject and message contents of unsolicited emails to individual users. [9] As the personalization of unsolicited messages continues to grow, the line between unsolicited bulk emails and good emails from known senders becomes less clear.

What is "graymail" and what are some proposed methods for detecting it?
While much of the emphasis of modern research has been placed on detecting messages traditionally defined as "spam" (i.e. messages that the end user did not sign up for), little emphasis has been placed on the detection of "graymail"—messages that could reasonably be considered either spam or good by different email users. [18-23] An analysis of 2.6 million email messages and labels showed that graymail accounts for approximately 25.4% of email messages. [18] Graymail is traditionally challenging to detect due to lack of labeled graymail data, as well as the unreliability of the properly labeled data that is available. [19] A study by Yih, *et al.* [20] concludes that, without taking into account a users' specific preferences, even an optimal filter will inevitably perform unsatisfactorily on graymail. Many proposed graymail detection methods seek to discover these user preferences by explicitly asking users to individually report the messages they believe to be graymail [22], yet such methods may pose scalability issues due to the added annoyance to the user of having to provide this feedback.

3. METHODOLOGY

To better assign human-defined classifiers to emails as organically as possible, our procedure aimed to combine researcher-guided email selections with an online survey. Due to its popularity and API availability, we chose to implement out survey with Google Mail (abbv. Gmail). The survey has 3 main sections: a first set in which we collected participant demographics, a second set of generic questions about their perception of spam and Gmail usage, and a third set where we asked detailed questions about a stratified sample of emails that each participant had in their actual Gmail account inbox.

3.1 Data Collection and Ethics

A key part to our survey was that participants were made to look at their own personal emails and answer questions based on them. To achieve this, we first had to have them read through our consent form (see Appendix E) and understand that, though they would be answering questions on the emails, the emails themselves would not be viewed or recorded by us. After participants consented, we had a researcher ask them how many emails they had in their inbox and use a random number selector to select ones for the participant to answer questions on.

Throughout the process, our primary concern was to maintain the participants' privacy and collect our data in an ethical manner. Towards this end, we took multiple steps to ensure their security:

1. We never saw or recorded participants' emails
2. We never collected participants' Gmail login credentials
3. We anonymized their responses and retained no individually-identifying information
4. If not on their personal machine, participants were asked to logout of their Gmail account at the end of the survey

3.2 Recruitment and Inclusion Criteria

We recruited participants from personal contacts and The University of Chicago campus. We limited participants to North America and also required them to be 18+. As our study's goal was to investigate common classifiers for spam among a wide range of emails, users were also subjected to another set of criteria based upon their Gmail account's metadata:

- More than 10 emails in the inbox and spam label
- An email account that was at least a week old

These criteria ensured that participants' accounts were sufficiently full enough to ask our detailed questions on a range of emails.

3.3 Email Selection

In our pilot study, we asked participants to look at 10 different emails and answer questions based on them. The base criteria for each of the emails were that they be 10 different emails from 10 different senders. The emails were selected randomly by a researcher using a random number generator and asking the participant to "open the *n*th email down in your ___". If the selected email failed one of our criteria (e.g, from the same sender), a new email would be selected. To give a more interesting spread, we had 5 emails be from the participant's inbox and the other 5 be from their spam label.

In future studies, it would be interesting to select emails on even more stringent criteria like their age and number of characters or attachments. This is something that should be handled programmatically. See Section 6: Moving Forward for more information on this survey automation.

3.4 Survey Structure

Our survey consisted of three main sections. We first asked participants about their demographics. We only collected basic demographics about participants, including age, gender, and profession. In the next section, we asked participants about their email usage and general characteristics of their account: how long have they had it, how many emails does it typically receive, purpose of the email account, etc.

The final section consisted of email-specific questions. First, we asked them to look at a randomly selected email within our email selection constraints. This was then followed by a set of questions that centered around whether the participant considered the email to be spam. For each email, we ask users to rank how confident they are that this message is spam, whether they signed up to receive this email, how often they receive emails from this sender, how likely they are to read the email, and whether they would like to continue receiving these emails, and how often they open and read similar messages. For each Likert question, a randomly selected half of participants will have their answers choices inverted. In addition to these questions, we ask users to categorize the message in one of several types of common email messages, including Unsolicited Marketing, Subscription Marketing, Daily/Weekly Newsletter, etc. We repeat this process for 12 emails. We have included a more detailed survey overview as well as our full survey instrument in our appendix.

After users have completed the survey, they are informed that they have completed the study and provide them with options for opting out of the study/having their data removed.

3.5 Analysis (Quantitative/Qualitative)

We plan to use the data collected by our survey to develop weighted classifiers on spam identification. By building a multi-factor regression logit model, we would be able to ascertain how likely an email category causes a user to consider it as spam. We will also treat the answers to the introductory questions as customizable thresholds for each individual email user. Those who are less strict on spam will use a model that is more forgiving of spam identifiers, while those who are stricter on spam will use a model that has a

lower threshold for spam identification. This would allow for customization of specific users.

3.6 Limitations

Because we will not have access to the contents of the emails before choosing them in the survey, it is possible that the random selection of user emails could misrepresent emails in general. This would also be skewed from the selection of participants as well. In future iterations of this project, a larger pool of participants should be created to help counter any skew from these uncontrollable potential confounds.

4. PILOT RESULTS

To better understand the general opinion on spam emails and how users classify them as well as identify specific concerns that participants had, we conducted a pilot test in a cognitive walkthrough survey format. Due to issues with building an interactive Flask and Qualtrics interface, we had participants open their emails manually using a random number generator to randomly pick emails and refrain from choosing emails with the same sender. We had 6 participants for our pilot survey.

Before conducting the pilot survey, we used a prewritten set of survey questions that consisted of both introductory/demographic and email specific questions. While the original set of questions included a question that asked them to categorize their email from a list of premade categories such as "email from school" and "marketing email", that particular question in the pilot survey was changed to only have an open text field box entry. By letting the participants categorize their emails in their own words, we were able to check if our original answer choices were comprehensive enough. While most of the participants' responses were already covered in one of our original categories, some could not be included in them such as "Facebook friend request update" and "Amazon delivery update". We used qualitative coding on the various uncovered descriptions by the participants to add more categories to the original answer choices. These categories added were "Social media related" and "Order status updates", which covered all of the originally uncovered email descriptions.

Participants also noticed redundant question choices from two different questions, one solely asking if the participant subscribed to the email and the categorize question that included both "unsolicited school email" and "solicited school email". We removed all answer choices that distinguished between solicited and unsolicited emails.

On the question that asked the participant to choose how confident they thought the particular email was spam, many asked what we meant by spam. Because we are building a spam filter that would decide which emails make it to their inbox, that question was changed to whether they wanted that email in their inbox with Likert Scale answer choices. The question was also used to determine the binary output variable in our model.

Due to the location that we used to recruit pilot study participants, all of the participants used their UChicago student emails. In order to make the study deployable to a broader audience, we added a question that asks the participant which type of email they are using (personal, school, work, etc). By including this metric, the model would be able to add another layer of possible classifications. We also added a question for the specific emails regarding how much of the total email they read.

During the survey, many participants asked the researcher present how long the whole survey would take. We added a brief description at the very beginning of the survey that included the estimated time of completion of 10-15 minutes.

Aside from the more prominent changes, we also changed a few ambiguities in the wording of the survey questions and answers from concerns by the participants, such as a more clear difference between "some college" and "4 year college" by editing the former to "some college no degree".

5. DATA ANALYSIS

The heading of a section should be in Times New Roman 12-point bold in all-capitals flush left with an additional 6-points of white space above the section head. Sections and subsequent sub-sections should be numbered and flush left. For a section head and a subsection head together (such as Section 3 and Subsection 3.1 of this document), use no additional space above the subsection head.

5.1 Approach to data analysis

We built a binary classifier that predicts whether a given email will be considered as spam based on its properties. The classifier is a logistic regression model, trained with Python's scikit-learn module using a certain proportion of our survey responses. The remaining proportion was a testing set used to evaluate the model.

5.2 Data Cleaning

We surveyed 7 students, and for each student, obtained responses for 10 emails, for a total of 70 emails. We made the simplifying assumption that the responses from each email is independent from each other. We converted to binary variable the results of "How confident are you that this message is spam". Values between 1 to 3 were coded as 0, and values from 5 to 8 were coded as 1. The value of 4 (neutral) was removed, leading us to take out 4 participants from the data analysis sample.

Some users did not indicate responses for some questions. As we want data point to be independent, or have the same number of samples, we removed the set of data from an individual who did not complete all questions. Two people's responses were removed because of this, bringing the total sample size to 52. Furthermore, we removed the first response because it was actually a trial attempt by one of our team members. Due to a confusion in initial communication, it was included in the analytical dataset. We have since removed this data point.

5.3 Description of variables

- *spam_result* - output of binary classifier (1=spam; 0=not spam)
- *sign_up* - did user sign up to receive this email
- *freq* - how often the user received this email from this sender
- *perc_spam* - percentage of emails user thinks this sender sends is spam
- *likely_read* - how likely to read emails from this sender
- *categorize_email* - removed due to lack of data points
- *continue_receiving* - would you like to continue receiving these emails
- *often_similar* - how often do you open and read emails like this one

5.4 Estimation of parameters

Our fitted model is a logistic regression model that expresses the logit-transformed probability of a binary outcome as a linear sum of predictor variables. Suppose that our binary spam outcome of (0,1) is expressed as y, and the probability p refers to the probability that y is equal to 1, in other words, an email is considered as spam. Let x1, x2 and so on refer to the predictor variables. The logistic regression model estimates parameters β0, β1 and so on as follows:

$$\textbf{logit(p)} = \textbf{log(p/(1-p))} = \beta 0 + \beta 1 * x1 + \ldots + \beta k * xk$$

Our model has multiple explanatory variables without interaction terms. Contextualizing the formula above to our experiment with our fitted parameters, we obtain the equation for our logit(p):

$$\textbf{logit(p)} = \beta 0 + \beta 1 * \text{sign\_up} + \beta 2 * \text{freq} + \beta 3 * \text{perc\_spam} + \beta 4 * \text{likely\_read} + \beta 5 * \text{continue\_receiving} + \beta 6 * \text{often\_similar}$$

β0	0.9172
β1	-0.1155
β2	0.9011
β3	-0.2150
β4	-0.4847
β5	0.5199
β6	-0.1504

Table 5.4: Estimated parameter values

Transcribing the numerical coefficients (Table 5.4), we obtain:

$$\textbf{logit(p)} = \textbf{0.917} + \textbf{(-0.1155)} * \text{sign\_up} + \textbf{(0.901)} * \text{freq} + \textbf{(-0.215)} * \text{perc\_spam} + \textbf{(-0.484)} * \text{likely\_read} + \textbf{(0.519)} * \text{continue\_receiving} + \textbf{(-0.150)} * \text{often\_similar}$$

5.5 Error analysis

We consider the model's accuracy through an error matrix, as well as its values of precision, recall, and F-score. An example of an error matrix is presented below (Figure 5.5.1). "TP" and "TN" represent "True Positive" and "True Negative" respectively. "FP" and "FN" represent "False Positive" and "False Negative" respectively.

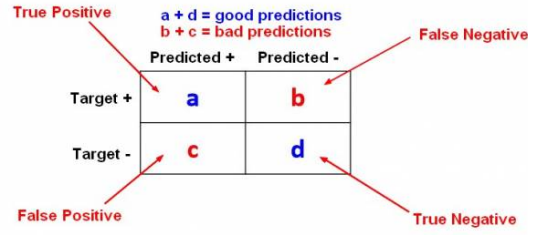

Figure 5.5.1: Example of an error matrix (accessed from the Gerardnico website)

As evident in our error matrix (Figure 5.5.2), we have 12 true values (7 + 5 = 12) and 1 false value (1 + 0 = 1). This suggests that our model was learned well from the training data.

7	0
1	5

Figure 5.5.2: Our error (confusion) matrix

We use an in-built Python function called "classification_report()" to look at the precision and recall metrics. We also consider the F1-score, which is a harmonic mean of the two (Table 5.5.3). These values all exceed 0.80, which look optimistic.

	Precision	Recall	F1-score	Support
0	0.86	0.86	0.86	7
1	0.83	0.83	0.83	6
Avg/Total	0.85	0.85	0.85	13

Table 5.5.3: Classification report of our model showing its precision, recall and F-score

5.6 Interpretation of results

In this section, we will interpret the odds ratio for each of the seven explanatory variables in our logistic regression model. Each explanatory variable may be interpreted as a "latent effect" or feature of a given email that affects the perception of that email as spam.

5.6.1 Relationship between spam perception and whether the user previously signed up to receive emails from this sender

The expected change in the odds of an email being spam for a unit increase in whether the user previously signed up to receive emails from the same sender, holding all other explanatory variables constant, would be $e^{(-0.1155)} = 0.891$. This means that if a user previously signed up to receive emails from the sender, then the odds are lower by a factor of 0.891 that he or she considers an additional email from this sender as spam.

5.6.2 Relationship between spam perception and frequency of past emails from same sender

The expected change in the odds of an email being spam for a unit increase in the frequency the user received an email for the same sender, holding all other explanatory variables constant, would be $e^{0.901} = 2.46$. In terms of intuition, this means that the more often a user receives an email from a particular sender, the higher the odds that the user would consider another email from this sender as spam.

5.6.3 Relationship between spam perception and whether the user thinks that most emails from this sender constitute as spam

The expected change in the odds of an email being spam for a unit increase in whether the user considers other emails from the same sender as spam would be $e^{(-0.215)} = 0.8065$. If a user previously signed up to receive emails from the sender, then the odds are lower by a factor of 0.8065 that he or she considers an additional email from this sender as spam.

However, this result is intuitively not sound. Our methodology in data analysis was valid, as we made sure to match the direction of training the data with the direction of the Likert scales. As a further next step, we would be curious to investigate the reason why when users think emails from the same user are spam, then paradoxically they are less likely to consider an email from this user as spam.

5.6.4 Spam perception vs Likelihood of reading the email
The expected change in the odds of an email being spam for a unit increase in whether the user is likely to read that email is $e^{(-0.485)} = 0.615$. If a user is likely to read that email, then it makes sense that he or she would not consider it as spam. The factor is 0.615 per unit increase in likelihood of reading that email. This finding confirms the intuition that one of the defining traits of spam is that users would not open an email they consider as so.

5.6.5 Spam perception vs Desire to continue receiving
The expected change in the odds of an email being spam for a unit increase in whether the user wants to continue getting emails from that sender would be $e^{0.520} = 1.68$. If a user has a strong desire to continue receiving emails, then it is more likely the user considers the email as spam. This is, again, paradoxical to general intuition. It is to be further investigated whether there is a latent factor beneath "continue receiving emails" that causes users to consider it not as spam; for example, if a user always receives an email from "University of Chicago Security Alert", while the email comes very often, because it is useful, the user wants to continue receiving it. However, the user may still classify it as "spam" in his or her mental model, because of its frequency of sending.

5.6.6 Spam perception vs Frequency of opening similar emails
The expected change in the odds of an email being spam for a unit increase in whether the user has opened similar emails is $e^{(-0.150)} = 0.861$. If a user has opened similar emails, then it is less likely the email is perceived to be spam by a factor of 0.861 per unit increase in the frequency of opening similar emails. This aligns well with the general intuition.

5.7 Data analysis programming code
The documentation in our Appendix shows the code that was used to implement the training and testing of the data using Python's scikit-learn.

6. MOVING FORWARD
For future, larger-scale iterations of this study that would aim to produce statistically significant results, we recommend several points of improvement and expansion.

6.1 Survey Automation
An obvious bottleneck on our ability to deploy this survey on a large scale is the requirement for researchers to be present during the course of the study. In our pilot study, our presence was necessary as we directed users on which emails to select from their inbox and answer survey questions on. For future iterations of this study, we looked into removing this bottleneck by automating this rather simple process.

First, we aimed to construct a Python script which would utilize Google's public Gmail API to pull unique, random emails and display them to the user. And we accomplished this. At present, we have written a Python script which can securely grab 10 random, unique emails according to our criteria from a user's inbox. The code for this script is in this paper's Appendix and our publicly available GitHub repository which is also linked in the Appendix.

One of our greater concerns while implementing this was making sure to maintain the participants' privacy. Namely, we wanted to preserve the following integrities:

1. participant login credentials should never be directly recorded
2. researchers should never be able to view participants' emails
3. access to participant inboxes should not persist beyond the study

We addressed these concerns through usage of OAuth2 and secure design practices. OAuth2 authentication allowed us to obtain a temporary token which would give our app read-only access to the participant's inbox. The token is then voided and deleted upon survey completion. Applying our secure design, our Python script only stores Gmail message ids in a text file which is deleted upon survey completion. We opted to only temporarily store ids rather than the actual messages themselves so that we would *never* store the actual emails and their contents on our devices. The ids are unique numbers which allow us to make an API request that will pull the specific emails' information in a json format; from there, our script can parse the json and display relevant information to the participant without storing any persistent records of the emails shown.

The reasons that we did not utilize this script in our pilot study were two-fold:

1. Our app was not verified by Google
2. It did not fully abstract out the researchers

To get an app verified by Google as secure, it requires a lengthy developer authentication and application process that takes around 2 weeks. Time constraints and not realizing the need for Google verification until it was too late left us with not enough time to complete this process. For future iterations of this study, we recommend starting this authentication process early; it is necessary for assuring the participant of their privacy and the study's security. Also, just a Python script cannot be integrated with Qualtrics surveys. A Flask web server must also be constructed.

6.2 Flask Web Server Development
The final form of our widely-deployable survey takes the form of a Qualtrics survey which dynamically passes the participant between itself and a Flask web server. The survey is meant to ask the participant questions and gather their responses. The web server is meant to run our Python scripts to authenticate, request, and display the participant's Gmail mail to them. Figure 6.2.1 diagrams this dynamic "passing" of the survey participant.

In this iteration of our survey, a researcher would not be needed and the survey could be widely distributed on services like Amazon's Mechanical Turk. Unfortunately, we were unable to complete implementing this iteration. Instead, we will give an overview of the steps needed to accomplish this:

1. Create a Flask-Python file with API routes to our Python script

2. Deploy it on a server like Heroku, Google App Engine, AWS Elastic Beanstalk, etc…
3. Create paths between the Qualtrics survey and Flask web server

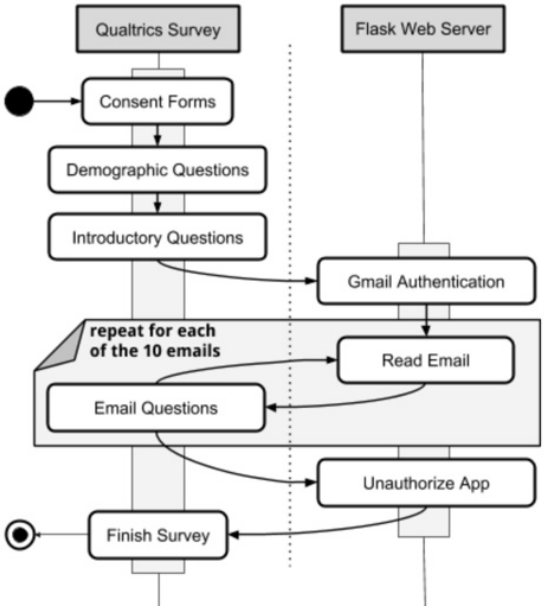

Figure 6.2.1: Merged activity diagram with sequence diagram that demonstrates which platform handles what actions for the participant in survey completion.

6.3 Collecting More Data Points

Aside from generating the results that our pilot survey does,t the study structure we have created can be expanded upon to gather other significant data with the hopes of generating other interesting results.

For example, future researchers may find it helpful to gather metadata about participants' inboxes and emails. Data points like the total amount of messages within the inbox, the actual senders of the emails, the amount of labels the participant uses, etc… all have the potential to be interesting statistics. To record this metadata, we recommend adding on an SQL Database to the Flask web server. One would also need to modify the ethics application to include this data.

7. CONCLUDING DISCUSSION

We draw out implications from our surveys and prior work.

Perceptions of Spam

In our original pilot study, we wanted participants to define what their definition of spam was. However, this was very difficult to many people and overall produced many different variations. From "emails I don't want to read" to "emails from senders I don't know", there was not a general consensus of how to define spam. The results of the study indicated that not even the individual participants were consistent with their definition of spam. Even for emails that they believed were spam, they were still likely to read them. This was especially the case for emails that were categorized as "Social media updates". Some types of emails that participants believed were not spam were very unlikely to be read. This was especially the case for newsletters. To many, this relationship is almost the opposite of what we would think of for spam.

Role of the Spam Filter

Because it is clear that people have different perceptions of spam, the optimal spam filter would be able to utilize the user's preferences in account when filtering out emails from the inbox. The role of the spam filter is meant to prevent certain emails from making it into the user's inbox, regardless of whether that user wants to open it or not. While the standard definition considers spam as unsolicited marketing to a broad audience, there are likely email users that would still prefer those emails in their inbox; whether or not they open and read it is up to them.

We hope that in the future, email spam filters will take into account various user preferences and their relationship to spam when deciding whether or not to let an email make it into the inbox. While it is harder to account for these preferences on the spot, the future of machine learning will eventually be able to identify past patterns and make better decisions on spam filtering.

8. ACKNOWLEDGEMENTS

We thank our CMSC 23210 instructors Blase Ur and Mainack Mondal, and our class TA's Weija He and Ahsan Pervaiz, for their insightful feedback and project review. Their guidance was pivotal in our production of this research pilot study.

We would also like to credit SOUPS for their assistance in providing a LaTeX and Word template for this paper.

References

[1] Jackson, Thomas, Ray Dawson, and Darren Wilson. "The cost of email interruption." *Journal of Systems and Information Technology* 5.1 (2001): 81-92.

[2] Caliendo, Marco, et al. "The cost impact of spam filters: Measuring the effect of information system technologies in organizations." (2008).

[3] Kraut, Robert E., et al. "Pricing electronic mail to solve the problem of spam." *Human-Computer Interaction* 20.1 (2005): 195-223.

[4] Marchewka, Jack T., Chang Liu, and Charles G. Peterson. "Perceptions of unsolicited electronic mail or Spam." *Journal of International Information Management* 12.1 (2003): 6.

[5] Bujang, Yanti Rosmunie, and Husnayati Hussin. "Should we be concerned with spam emails? A look at its impacts and implications." *Information and Communication Technology for the Muslim World (ICT4M), 2013 5th International Conference on*. IEEE, 2013.

[6] Rao, Justin M., and David H. Reiley. "The economics of spam." *Journal of Economic Perspectives* 26.3 (2012): 87-110.

[7] Fallows, Deborah. "Part 3. The Volume and Burdens of Spam." Pew Research Center: Internet, Science & Tech, 22 Oct. 2003, www.pewinternet.org/2003/10/22/part-3-the-volume-and-burdens-of-spam/.

[9] Gashti, Mehdi Zekriyapanah. "Detection of Spam Email by Combining Harmony Search Algorithm and Decision Tree." *Engineering, Technology & Applied Science Research* 7.3 (2017): 1713-1718.

[10] Ezpeleta, Enaitz, Urko Zurutuza, and José María Gómez Hidalgo. "A study of the personalization of spam content using Facebook public information." *Logic Journal of the IGPL* 25.1 (2016): 30-41.

[11] Salehi, Saber, et al. "Fuzzy granular classifier approach for spam detection." *Journal of Intelligent & Fuzzy Systems* 32.2 (2017): 1355-1363.

[12] da Silva, Luis Alexandre, et al. "Learning spam features using restricted boltzmann machines." *IADIS-INTERNATIONAL JOURNAL ON COMPUTER SCIENCE AND INFORMATION SYSTEMS* 11.1 (2016): 99-114.

[13] Song, Jonghyuk, Sangho Lee, and Jong Kim. "Spam filtering in twitter using sender-receiver relationship." *International Workshop on Recent Advances in Intrusion Detection*. Springer, Berlin, Heidelberg, 2011.

[14] Fumera, Giorgio, Ignazio Pillai, and Fabio Roli. "Spam filtering based on the analysis of text information embedded into images." *Journal of Machine Learning Research* 7.Dec (2006): 2699-2720.

[15] Saad, Omar, Ashraf Darwish, and Ramadan Faraj. "A survey of machine learning techniques for Spam filtering." *International Journal of Computer Science and Network Security (IJCSNS)* 12.2 (2012): 66.

[16] Christina, V., S. Karpagavalli, and G. Suganya. "A study on email spam filtering techniques." *International Journal of Computer Applications* 12.1 (2010): 0975-8887.

[17] Li, Jiwei, et al. "Towards a general rule for identifying deceptive opinion spam." *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1. 2014.

[18] Bhowmick, Alexy, and Shyamanta M. Hazarika. "Machine Learning for E-mail Spam Filtering: Review, Techniques and Trends." *arXiv preprint arXiv:1606.01042* (2016).

[19] Chang, Ming-wei, Scott Yih, and Robert McCann. "Personalized Spam Filtering for Gray Mail." *CEAS*. 2008.

[20] Sculley, D., and Gordon V. Cormack. "Filtering Email Spam in the Presence of Noisy User Feedback." *CEAS*. 2008.

[21] Yih, Wen-tau, Robert McCann, and Aleksander Kolcz. "Improving spam filtering by detecting gray mail." *CEAS-2007*(2007).

[22] Nigam, Paras, Mohammed Mohsin Dalla, and Dilip Kumar Gudimetla. "Graymail filtering-based on user preferences." U.S. Patent No. 9,954,805. 24 Apr. 2018.

[23] "Why Antispam Is Not Enough Anymore?" *CloudBizz*, www.cloudbizz.com/wp-content/uploads/2015/08/WP_Why_Antispam_is_not_enough_anymore.pdf.

[24] Isacenkova, Jelena, and Davide Balzarotti. "Shades of gray: A closer look at emails in the gray area." *Proceedings of the 9th ACM symposium on Information, computer and communications security*. ACM, 2014.

[25] Banday, M. Tariq, and Tariq R. Jan. "Effectiveness and limitations of statistical spam filters." *arXiv preprint arXiv:0910.2540* (2009).

APPENDIX

A. Detailed Survey Overview

Our survey consisted of three main sections. First, we asked participants about their demographics. We only collected basic demographics about participants, including age, gender, and profession. We have included a more detailed survey overview as well as our full survey instrument in our appendix. We then asked participants about their email usage and general characteristics of their account: how long have they had it, how many emails does it typically receive, purpose of the email account, etc.

The next section consisted of email-specific questions. First, we displayed information about the email selected for the participant (sender, receiver, message, attachments, etc.) This was then followed by a set of questions that centered around whether the participant considered the email to be spam. For each email, we ask users to decide if they wanted it in their inbox versus the spam folder, whether they signed up to receive this email, how often they receive emails from this sender, how likely they are to read the email, and whether they would like to continue receiving these emails, and how often they open and read similar messages. In addition to these questions, we ask users to categorize the message in one of several types of common email messages, including Marketing, School-related, Daily/Weekly Newsletter, etc. We repeat this process for 10 emails.

After users have completed the survey, they are informed that they have completed the study and provide them with options for opting out of the study/having their data removed.

B. Survey Forms

The study will seek to understand what users perceive as spam. To participate in this study, you will be required to provide read-access to your Gmail inbox. There will be two parts to this study. First, we will ask introductory questions about your general background and how you interact with spam. In the second part, you will be asked to answer questions about 10 randomly selected emails from your Gmail inbox.

Demographics

8.1 What is your gender?
- ○ Male
- ○ Female
- ○ Other (please specify)
- ○ Prefer not to answer

8.2 What is your age?
- ○ Under 18
- ○ 18 - 24
- ○ 25 - 34
- ○ 35 - 44
- ○ 45 - 54
- ○ 55 - 64
- ○ 65 +
- ○ Prefer not to answer

8.3 What is highest degree of education you have obtained? (If currently enrolled, highest degree received)
- ○ Some high school experience
- ○ High school graduate
- ○ Some college credit, no degree
- ○ 2 year college degree
- ○ 4 year college degree
- ○ Masters degree
- ○ Ph.D.
- ○ Prefer not to answer

8.4 For this study, what kind of email will you be using?
- ○ Personal
- ○ Work
- ○ School
- ○ Other

Introductory Questions

8.5 How often do you receive spam in your email inbox?
- ○ Less than once a week
- ○ Once a week
- ○ Twice a week
- ○ Every other day
- ○ Once a day
- ○ Twice a day
- ○ More than twice a day

8.6 How many minutes per day do you spend dealing with spam?
- ○ Less than one minute
- ○ 1 - 5 minutes
- ○ 6- 10 minutes
- ○ 11+ minutes

8.7 Approximately what percentage of your email is unsolicited?
- ○ Less than 5%
- ○ 5 - 25%
- ○ 26 - 50%
- ○ 51 - 75%
- ○ 75 - 100%

8.8 I am satisfied with the Gmail spam filter.
- ○ Strongly Disagree
- ○ Disagree
- ○ Somewhat Disagree
- ○ Neutral
- ○ Somewhat Agree
- ○ Agree
- ○ Strongly Agree

Email Specific Questions

8.9 Approximately what percentage of this email did you read when you received it?

- Less than 5%
- 5 - 25%
- 50 - 75%
- 75 - 95%
- More than 95%

8.10 Are you familiar with the sender of this email?

- Yes
- No

8.11 Did you sign up to receive this email?

- Yes
- No

8.12 How often do you receive emails from this sender?

- Less than once a week
- Once a week
- Twice a week
- Every other day
- Once a day
- Twice a day
- More than twice a day

8.13 Approximately what percentage of emails from this sender belong in your inbox?

- Less than 5%
- 5 - 25%
- 25 - 50%
- 50 - 75%
- 75 - 95%
- More than 95%

8.14 Which of the following describe this email (check all that apply)?

- Marketing
- Job board posting
- Social media related
- Daily/Weekly Newsletter
- Order status update
- Personal email
- Work-related email
- School-related email
- Other(s) (please specify)

8.15 I am likely to read the contents of this email.

- Strongly Agree
- Agree
- Somewhat agree
- Neutral
- Somewhat disagree
- Disagree
- Strongly disagree

8.16 This email belongs in my inbox.

- Strongly Agree
- Agree
- Somewhat agree
- Somewhat disagree
- Disagree
- Strongly disagree

Do you consent to finishing the survey and allowing your answer to be used to better understand spam classification for future applications? If no, we will remove this survey from our data collection.

- Yes, I consent
- No, I do not consent

C. Code Scripts

This script can also be downloaded from our GitHub repository:
https://github.com/cchoy96/USP18-IdentifyingSpam

```
from __future__ import print_function
from apiclient.discovery import build
from httplib2 import Http
from oauth2client import file, client, tools
from random import randint
import os, json, base64


credFile = 'storedcredentials.json'
mailFile = 'storedEmail_ids.txt'


# Setup the Gmail API
def authenticate ():
    SCOPES = 'https://www.googleapis.com/auth/gmail.readonly'
    store = file.Storage(credFile)
    creds = store.get()
    if not creds or creds.invalid:
        flow = client.flow_from_clientsecrets('client_secret.json', SCOPES)
        creds = tools.run_flow(flow, store)
    gmail_service = build('gmail', 'v1', http=creds.authorize(Http()))
    return gmail_service


# Make a GET query to retrieve a msg based on its id
def getMsg (GMAIL, m_id):
    return GMAIL.users().messages().get(userId='me', id=m_id).execute()


# Print relevant information from the message json
def printMsgDetails (message):
    hdrs = message['payload']['headers']
    subj = sender = target = cc = '[none]'
    for entry in hdrs:
        key = entry['name'].encode('utf-8')
        val = entry['value'].encode('utf-8')
        if key == 'Subject':
            subj = val
        if key == 'From':
            sender = val
        if key == 'To':
            target = val
        if key == 'CC':
            cc = val
    snippet = message['snippet'].encode('utf-8')
    print('Subject: {s}\nFrom: {f}\nTo: {t}\tCC: {c}\n{snip}\n=========\n'
          .format (s=subj, f=sender, t=target, c=cc, snip=snippet))


# Call the Gmail API
def getMessages (GMAIL, query):
    result = GMAIL.users().messages().list(userId='me', q=query).execute()
    msgs = result.get('messages', [])
    if not msgs:
        print('\tNo msgs found. Exiting...')
        exit(1)


    return msgs
```

```
# Extract n random, unique emails from the given set of emails
def extractRandomUnique (GMAIL, file, msgs, n):
    id_list = []
    for m in msgs:
        id_list.append(m['id'])
    i = 0
    max_idx = len(id_list) - 1
    while i < n:
        surveyMail = []
        surveyMail_ids = []
        surveyMail_snd = []
        for line in file:
            surveyMail.append(json.loads(line))
        for mail in surveyMail:
            surveyMail_ids.append(mail['id'])
        for mail in surveyMail:
            for entry in mail['payload']['headers']:
                if entry['name'].encode('utf-8') == 'Sender':
                    surveyMail_snd.append(entry['value']).encode('utf-8')
                    break


        # Randomize Email Selection. Don't select same email twice
        rand_idx = randint(0, max_idx)
        if id_list[rand_idx] in surveyMail_ids:
            print ('Duplicate mail id!')
            continue


        # Filter out duplicate senders
        target_msg = getMsg(GMAIL, id_list[rand_idx])
        for entry in target_msg['payload']['headers']:
            if entry['name'] == 'Sender':
                if entry['value'] in surveyMail_snd:
                    print('Duplicate Sender!')
                    continue
                else:
                    break


        file.write(str(id_list[rand_idx]) + '\n')
        i += 1


# Print out the email at IDX line in FILE
def printSurveyMail (GMAIL, file, idx):
    f = open(file, 'r')
    ids = []
    for line in f:
   ids.append(line[:-1])
    printMsgDetails(getMsg(GMAIL, ids[idx]))
    f.close()


# Remove Stored Credentials
def cleanup ():
    os.remove(credFile)
    os.remove(mailFile)
    print('Removed stored credentials and emails!')


def main ():
    print('Authenticating...')
    serv = authenticate()
    print('Authentication Complete!')
```

```
    print('Building 10 Survey Emails...')
    inboxMail = getMessages(serv, 'is:inbox')
    spamMail = getMessages(serv, 'is:spam')


    f = open(mailFile, 'w+')
    extractRandomUnique(serv, f, inboxMail, 5)
    extractRandomUnique(serv, f, spamMail, 5)
    f.close()


    print('Printing Emails...\n')
    for i in range(0,10):
        printSurveyMail(serv, mailFile, i)


    cleanup()
    print('\n---------\nScript Complete!')

main ()
```

This was the code used for our data analysis:

```
import pandas as pd
from sklearn import model_selection
from sklearn import linear_model
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report

# Convert ordinal output variable (1,2,3,5,6,7) into binary output variable (0,1)
# There were three responses that had "4" (neutral) as output variable
# We removed these responses from our data analysis sample

def convert_outcome_to_binary(df):
   num_of_responses = len(responses) - 1
   i = 0
   while i < num_of_responses:
       if responses.iloc[i][1] > 4: # binary outcome 1 = yes to spam
           responses.at[i,'1_Q8'] = 1;
       elif responses.iloc[i][1] < 4: # binary outcome 0 = no to spam
           responses.at[i,'1_Q8'] = 0;
       else:
           raise Exception('Reached the value 4 for supposedly binary classifier')
       i = i + 1
   print(responses)

def train_predict_evaluate(df):
   # First we train the model
   X = df.iloc[:, 2:9]
   y = df.iloc[:, 1:2]
   print(X)
   print(y)
   X_train, X_test, y_train, y_test = model_selection.train_test_split(X, y)
   logit = linear_model.LogisticRegression()
   logit.fit(X_train, y_train)

   # Second we predict the test value
   y_predicted = logit.predict(X_test)

   # Third we evaluate the classifier using a confusion, or error, matrix
   # This is done between the test binary output and the predicted binary output
```

```
    error_matrix = confusion_matrix(y_test, y_predicted)
    print(error_matrix)

    # Fourth we look at the F-value, which combines metrics of precision and recall
    # Precision and recall are measures of false positives and false negatives
    print(classification_report(y_test, y_predicted))

    # Find the coefficients of the logistic model
    coef = logit.coef_ # clarify this
    print coef

if __name__ == '__main__':
    responses = pd.read_csv('spamdata_cleaned.csv')
    convert_outcome_to_binary(responses)
    train_predict_evaluate(responses)
```

D. Ethics Application

Usable Security and Privacy, Spring 2018

20 April 2018

1) Study title
Machine Learning: Defining and Identifying Spam

2) Names of investigators (i.e., students in your group)
Christopher Choy, Bijan Oviedo, Scott Wang, Steven Wei
{cchoy, boviedo, zirenw, srwei} @uchicago.edu

3) Study location. Explain where the research activities will take place (including recruitment, data collection, and/or data analysis.)

Recruitment will take place in person, where we will attempt to recruit ~12 participants by posting to UChicago Facebook groups online. Study will take place online, and participants will not need to be in the same location as researchers. Data analysis will take place at various buildings on UChicago's campus.

4) Will any of your research procedures occur outside the United States?
[] Yes
[x] No

5) Provide a brief, non-technical description of the purpose of the research, including the research question(s) you hope to answer.

By understanding what users perceive as spam, one can improve machine learning classifiers designed to detect spam.

What kinds of specific markers do users personally employ to identify spam?
How does a spam detector using these specific markers compare to current ones in the market? (Gmail Spam)

6) Which research procedures does this study involve? (Check all that apply)
[x] Surveys / Questionnaires
[] Interviews / Focus groups
[] Observational / Ethnographic research
[] Secondary data analysis (analysis of data that already exists)
[] Audio or video recording or photographs
[] Deception / incomplete disclosure of research purpose or procedures
[] Other

7) In non-technical language, describe the procedures subjects will be asked to complete or undergo. Explain step by step what subjects will be asked to do. If your study includes multiple variations of the procedures, please make clear the procedures included in the variations.

Subjects will be asked to perform an online survey where they will allow us access to their google email inbox to pull their individual emails. The questions will ask them for non-identifiable information about the emails (such as if it is from someone in their immediate family) and how "spammy" they consider the email.

----Participants and Recruitment------

8) Approximately how many participants do you anticipate enrolling in this study (at all research locations / sites)?

We aim to have 12-16 participants involved in our study. More would be preferred for greater statistical significance, but the goal of this research is primarily to develop a scalable infrastructure for our study.

9) Describe the criteria for enrollment -- will you be limiting your enrollment to a certain age range, gender, people with certain health conditions, etc.? Please also describe any factors that will exclude people from enrollment.

We will not be limiting our enrollment to individuals in certain population demographics; however, we will only use participants who consent to giving the researchers temporary access to their inbox through the Gmail API, which may create bias in the demographics recruited.

10) Vulnerable populations -- check the boxes for ALL vulnerable populations from which you may enroll participants:
[] Children
[] Wards of the state
[] Prisoners / detainees
[] Adults not competent to consent
[x] Employees or students of the University of Chicago
[] Non-English speakers
[] Other vulnerable populations

11) Who will be recruiting individuals for participation in this research project? Explain whether it will only be members of the University of Chicago research team, collaborating researchers at other institutions, or others (e.g., a survey firm hired by the research team) who will be doing the recruitment activities.

Recruiting will be done entirely by the primary four student investigators of our research group.

12) Please check off all methods of recruitment that will be used:
[x] Directly approaching participants (in-person recruitment)
[] Email / listserv / electronic mailing list
[x] Flyers / posters or brochures
[] Letters sent to potential participants
[] Radio / television / video announcements
[] Newspaper / magazine advertisements
[x] Website / social media posting such as Craigslist, Facebook, UChicago Marketplace, etc.
[] Telephone scripts
[] Amazon Mechanical Turk
[] SONA system
[] Snowball sampling
[] Other

13) Provide details on your recruitment methods, including names of any publications / websites in which you will post recruitment information.

Recruitment will be primarily done by approaching people and asking if they would like to participate in our study. Further recruitment may be done by posting a link to our survey on social media like Facebook.

14) Attach all recruitment script, flyers, social media postings, and other materials you plan to use for recruitment purposes.

Social Media Posting (FB):
"Hey guys, if you have the time and wouldn't mind helping me out with my research project for the quarter, spare me ~30min and fill out this survey! If you complete it, you can enter into a raffle for a $50 Amazon gift card! The survey centers itself around answering the question: *How do users perceive and define spam not caught by conventional spam filters?* "

15) Will your study offer any compensation / incentive to research participants (including cash, gift cards, course credit, buying the participant a meal, etc.)

We will compensate participants by allowing them to opt-in to a raffle for a $50 Amazon gift card.

----Risks------

16) Describe the forseeable risks associated with your study. Please include discussion of any non-physical risks, such as economic, psychological, social, and legal harms.

There may be a few minimal risks posed to participants. In particular, participants may be at risk to suffer some psychological/social harm as they may be shown an unflattering email from their inbox or spam filters. They may believe that researchers have access to these emails (we won't) or feel uncomfortable answering questions about these emails.

17) Describe the steps you will take to minimize risks to your participants (for example, using pseudonyms or a coding system, etc.)

Risk to participants will be minimized by anonymizing their data and responses. We will also use third-party services like OCaml 2.0 to ensure privacy of their email login information. We will also not record any data from their emails. All recorded data will come from their voluntarily provided responses. At any point during the survey, participants may opt out and have their data removed.

18) If applicable to your study, what steps will you take if a participant becomes distressed during your study or reports intent to harm themselves or others?

If a participant becomes distressed during our study, we will allow them to stop the survey immediately and have their responses/data omitted from our results. So long as a discernible effort to complete the survey was given, participants will still be allowed to opt into the compensation raffle even if their data was not recorded.

If a participant reports intent to harm themselves or others, we will ask them to cease the survey immediately and remove their responses from our records. They will not be offered compensation. This honestly, isn't all that applicable to our research project; we're just asking them to answer some questions about spam.

----Data Collection and Protection------

19) In what format will the research data be collected and stored?
[] Paper
[x] Electronic
[] Audiovisual / recording media
[] Stored biological specimens
[] Artifacts
[] Other

20) Explain where the research data will be stored while the study is active (e.g., UChicago Box, personal laptop, thumb drive, departmental computer server, office file cabinet, etc.)

Research data will be stored on Google Form's servers, where participants will record their responses.

21) What security measures will be in place for each type of data to minimize the possibility of a data breach (password protection, encryption, locked file cabinet in a locked office, behind a firewall, etc.)

All data from the survey will be stored in a database that is password protected. Access and authentication for accessing their emails will be done through OCaml 2.0 authentication.

22) Will you collect any identifiers from the research participants (including names, addresses, Social Security Numbers, email and phone contact information, etc.)?

In order to use the Gmail API, we will collect their names and their email addresses. This is also primarily to ascertain we do not get repeat responses from people. This data will be stored separately from their responses.

23) What identifying information about research participants will be linked to the data?
[] Data will be directly labeled with personal identifying information
[x] Data will be labeled with a code that the research team can link to personal identifying information through a crosswalk to the coding system
[] Data will be labeled with a code but the research team will not have access to the crosswalk that connects codes to participant identifiers
[] Data will not be labeled with any identifying information and a coding system will not be used
[] Other

24) If you will be using a coding system, who will have access to the crosswalk that links participant identifiers to the data, and where will you store the crosswalk?

N/A

----Consent------

25) Check which type of consent process you plan to use with adult participants (select all that apply):
[] Written consent form signed by the participant
[x] Information sheet / consent script without participant's signature (if using a verbal consent process or online consent script)
[] Request to alter consent (some elements of consent waived)
[] Request to waive consent -- consent not being obtained
[] Not applicable -- no adults will be enrolled as research participants

26) Who will obtain consent from participants? Will the Principal Investigator, other members of the University of Chicago research team, collaborating researchers from other institutions, or another third party (such as a survey firm) obtain consent?

Researchers will obtain consent from the participants via an online consent form process that is bundled with our survey.

27) Describe the process that will be used to obtain consent, including how, when, and where consent will be discussed. If you might enroll any illiterate individuals, please explain how you will obtain consent from those individuals.

Consent will be obtained at the beginning of the survey. In a description, we will explain in broad terms the purpose of this study, inform them that they will need to provide read-access to their Gmail inbox, and that the researchers will only store metadata about the email (both computed metadata as well as metadata provided by them), but will never store the actual text of the emails. Participants will have the option to "opt out" of providing information about any specific email for any reason, and this will simply move the survey on to the next email they will be asked to provide information about. At the end of the survey, we will give users the option to review their provided answers and withdraw their responses either from any individual email or from the study as a whole.

----UChicago Affiliates------

28) (If enrolling UChicago students or employees) Explain how you will minimize the potential for employees and/or students of the University of Chicago to feel coerced to participate in the research.

All potential subjects will be fully briefed on the purpose of the survey as well as the possible harmful risks associated with it, including their consent to gain access to their email inboxes. They will also be given the option to opt out of the survey during and after its completion.

----Surveys------

29) Describe all surveys / questionnaires to be used in this study

Our survey will have three main parts. The first part will consist of informing the participant about the survey process and obtaining their consent and authentication to access their emails through the GMail API. The second part will consist of showing participants emails from their inbox and asking them to answer spam-related questions about the email. The third section of the survey will consist of gathering demographic information and survey feedback.

30) How often will participants be asked to complete the surveys / questionnaires and approximately how long will it take to complete the surveys / questionnaires?

Participants will only be asked to complete the survey once. The estimated time of completion is 30 minutes.

31) Will you be using any survey software?
[x] Yes
[] No

32) Attach the full text of any surveys / questionnaires you plan to use.
See above.

----Deception------

38) Describe what information will be withheld from participants or what misinformation will be provided to participants.

N/A. We will not be withholding information from participants.

39) Explain why this research involves no more than minimal risk to participants and why it would be impracticable to carry out the research without the use of deception/incomplete disclosure.

This research poses no more than minimal risk to the participants as we will only store data about them that they voluntarily gave us and the data will be anonymized. Regarding access to their personal emails, usage of OCaml 2.0 authentication will ascertain that we cannot store or misuse their login credentials. Our study will not make use of deception of incomplete disclosure.

40) Describe the plans for debriefing participants after their participation. If you do not plan to debrief participants, explain why.

n/a - We will not be debriefing them as we do not plan to use deception or incomplete disclosure tactics.

41) Attach the full text of any debriefing script/statement that you will use.

n/a.

----Additional attached documents------

42) Attach the full text of consent forms. See the following model consent forms:
See Appendix E.

43) Attach any additional study materials not previously requested
N/A

E. Consent Form

UNIVERSITY OF CHICAGO
CONSENT FORM FOR RESEARCH PARTICIPATION

Study Title: Machine Learning: Defining and Identifying Spam

Principal Investigator:

Student Researcher: Christopher Choy, Bijan Oviedo, Scott Wang, Steven Wei

IRB Study Number: _____

We are a group of students at the University of Chicago in the Department of Computer Science. We are planning to conduct a research study, which we invite you to take part in. This form has important information about the reason for doing this study, what we will ask you to do if you decide to be in this study, and the way we would like to use information about you if you choose to be in the study.

Why are you doing this study?
You are being asked to participate in a research study about using machine learning to define and identify spam. The purpose of the study is to improve machine learning classifiers designed to detect spam by understanding what users perceive as spam.

What will I do if I choose to be in this study?
You will be asked to perform an online survey where you give us access to your Google email inbox to pull emails. The questions will ask you for non-identifiable information about the emails (such as if it is from someone in the immediate family) and how "spammy" you would consider the email..

Study time: Study participation will take approximately 30 minutes in a single session. This is the total time commitment.

Study location: All study procedures will take place online, but the researcher analysis and recruitment process may take place at specific physical locations in the University campus.

What are the possible risks or discomforts?

Your participation in this study does not involve any physical or emotional risk to you beyond that of everyday life. As with all research, there is a chance that confidentiality of the information we collect from you could be breached – we will take steps to minimize this risk, as discussed in more detail below in this form.

What are the possible benefits for me or others?
You are not likely to have any direct benefit from being in this research study. This study is designed to learn more about detecting spam emails. The study results may be used to help other people in the future.

How will you protect the information you collect about me, and how will that information be shared?
Results of this study may be used in publications and presentations. Your study data will be handled as confidentially as possible. If results of this study are published or presented, individual names and other personally identifiable information will not be used.

To minimize the risks to confidentiality, we will not have access to view the contents of your emails and will store the metadata in a password protected database.

We may share the data we collect from you for use in future research studies or with other researchers – if we share the data that we collect about you, we will remove any information that could identify you before we share it.

If we think that you intend to harm yourself or others, we will notify the appropriate people with this information.

Financial Information
Participation in this study will involve no cost to you. You will have the option to opt-in to a raffle for a $50 Amazon gift card.

What are my rights as a research participant?
Participation in this study is voluntary. You do not have to answer any question you do not want to answer. If at any time and for any reason, you would prefer not to participate in this study, please feel free not to. If at any time you would like to stop participating, please tell me. We can take a break, stop and continue at a later date, or stop altogether. You may withdraw from this study at any time, and you will not be penalized in any way for deciding to stop participation.
If you decide to withdraw from this study, the researchers will ask you if the information already collected from you can be used.

[This section is required if U. of Chicago students are being recruited]
What if I am a University of Chicago student?
You may choose not to participate or to stop your participation in this research at any time. This will not affect your class standing or grades at University of Chicago.

[This section is required if U. of Chicago employees are being recruited]
What if I am a University of Chicago employee?

Your participation in this research is in no way a part of your university duties, and your refusal to participate will not in any way affect your employment with the university, or the benefits, privileges, or opportunities associated with your employment at University of Chicago.

Who can I contact if I have questions or concerns about this research study?
If you have questions, you are free to ask them now. If you have questions later, you may contact the researchers at

Christopher Choy: cchoy@uchicago.edu
Bijan Oviedo: boviedo@uchicago.edu
Scott Wang: zirenw@uchicago.edu
Steven Wei: srwei@uchicago.edu

If you have any questions about your rights as a participant in this research, you can contact the following office at the University of Chicago:

Social & Behavioral Sciences Institutional Review Board
University of Chicago
1155 E. 60thStreet, Room 418
Chicago, IL 60637
Phone: (773) 834-7835
Email: sbs-irb@uchicago.edu

Consent
I have read this form and the research study has been explained to me. I have been given the opportunity to ask questions and my questions have been answered. If I have additional questions, I have been told whom to contact. I agree to participate in the research study described above and will receive a copy of this consent form.

Consent for use of contact information to be contacted about participation in other studies
Initial one of the following to indicate your choice:
_____ (initial) I agree to allow the researchers to use my contact information collected during this study to contact me about participating in future research studies.
_____ (initial) I do not agree to allow the researchers to use my contact information collected during this study to contact me about participating in future research studies.

_____ Participant's Name
_____ Participant's Signature
_____Date