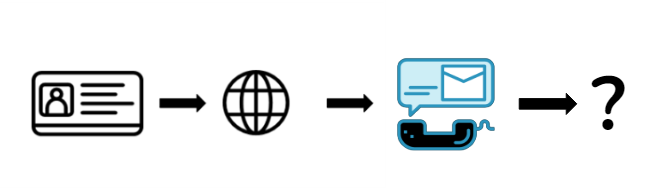
**First Interaction Experiment with Newsletters – Summer 2024**

Introduction

In the Summer of 2024, we proposed a research question that requires an additional step of interaction on top of a single one-time transaction we conduct during our sign-up event. Up until now, the Use & Abuse (U&A) team have developed a framework to conduct various research questions by tracking the movement of personal information across the web through a one-time online transaction. The framework requires the creation of fake identities (ID) using a pseudo random number generator (pRNG), conduct a sign-up in a semi-automated event, then subsequently collect and analyze all emails, SMS texts, and voicemails received to trace the mentions of the fake IDs across the internet. The one-time transaction in each fake ID allows traceability of how information is being spread, as each information is unique and tied to each fake ID.

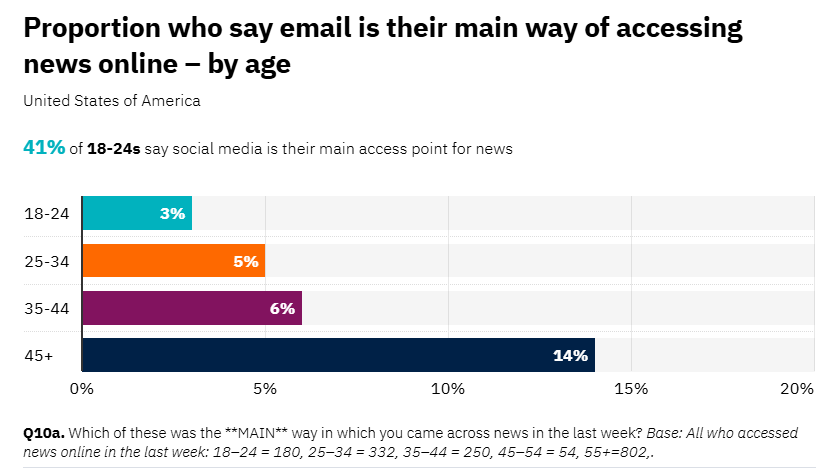
Figure 1. Displays the current framework of the Use & Abuse project.

As we continue to observe the complexity of the web through various research questions, we see many potentials with adding an interaction engine to the current framework. There has never been a focus on the interaction with emails, SMS texts, and voicemails, so through this experiment we aim to further the building of an interaction engine. By acknowledging the current framework of the U&A project, we built this as a smaller experiment to help bridge the gap between what we have and what we hope to achieve for a larger-scale experiment in the future. Due to the current system, we will be limiting our interactions for this experiment to just emails. For implementation, we want to add an additional step of interaction while we collect emails to understand the potential what interacting can do. Do organizations collect users’ clicks to send more relevant emails based on what was clicked? Are organizations spamming or targeting users once they start interacting? As we uncover various findings in this experiment, we will be able to further the development of the framework to tackle more complex research questions.

To understand the action of engaging with emails to build an interaction engine, we chose to interact with newsletters sent from media outlets. Media outlets hold a great amount of power with the platforms they use to disseminate information, but in doing so they also receive skepticism from individuals on the reliability of their works. With less newspapers and more newsletters, we aim to see how media outlets may potentially collect and use data based on the interaction (i.e., how many emails are being sent, promotional email, etc.). The initial presentation/proposal of the question can be found [here](https://virginiatech-my.sharepoint.com/:p:/g/personal/sharonkim213_vt_edu/EVeYDkYWUqVLpDY4tMfsBGYBOn2MHDRxLqfGmD8n-YXJAQ?e=LeIenT).

Background

Newsletters have opened opportunities for individuals and small publishers to publish content, as well as attract new customers and build loyalty among mainstream media outlets. A Digital News Report made by Reuters Institute shows that in the United States 22% of people use newsletters or email alerts. A distinguishing feature of email newsletters that appeal to people today, noted by Reuters Institute, is the personality of newsletters.

Figure 2. A chart created by Reuters Institute showing the percentage of people who say they access news online through email by age.

However, email newsletters remain popular among older individuals (like shown in figure 2) [[3],](https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2022/email-news-its-contribution-to-engagement-and-monetisation) while social media continues to grow to become very popular and meaningful among Gen Zs and Millennials [[4]](https://www2.deloitte.com/us/en/insights/industry/technology/media-industry-trends-2023.html). Why may social media be so popular among young individuals? Social media is run by computer programs that collect massive quantities of data to create content for its users [[5]](https://www.pewresearch.org/internet/2018/11/16/algorithms-in-action-the-content-people-see-on-social-media/). With that media outlets find themselves competing with social media platforms as many news outlets struggle to get clicks from readers. The voices in the online spaces like social media overshadow many news outlets [[6]](https://news.un.org/en/story/2022/03/1113702). This raise concerns as many social media platforms spread misinformation, fake news, and disinformation [[7]](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9910783/). As media outlets compete with social media platforms and grow in reaching more readers digitally, we aim to learn more about the newsletters of different media outlets (mainstream, secondary, and tertiary sites). After an initial analysis of CNN’s privacy policy, we have structured our question to ask how data usage and collection may affect the contents readers receive [[8]](https://virginiatech-my.sharepoint.com/:b:/g/personal/sharonkim213_vt_edu/EbE7u9szuAhNnSo9svwEJjkBpcOH3cvKA_KTOPd6r_DSqg?e=440qec). How are media outlets using data to reach more readers and create more personalized newsletters to engage readers? Through this experiment, we hope to see the interaction behaviors of newsletters and improve the current system of U&A.

Experimental Set-Up

To interact with the newsletters, the same framework from the Use & Abuse (U&A) project was kept with the only additional step of interaction after the sign-up event. The selection of media outlets was based on the chart created by Allsides (accessed on 25 July 2024) [[12](https://www.allsides.com/media-bias/media-bias-chart), [13](https://virginiatech-my.sharepoint.com/:b:/g/personal/sharonkim213_vt_edu/EaZOxzf_wwdEkEYj6XPbEOMBYD0VbLrXx9J2ZBufDbRJqw?e=vLxJAt)]. The chart categorized media outlets into five groups (left, lean left, neutral, lean right, right) and from each group two media outlets were chosen. Using the chart, the choice of media outlets for all groups except the left group was narrowed down based on the monthly visits reported by Statista[[1,](https://www.statista.com/statistics/381569/leading-news-and-media-sites-usa-by-share-of-visits/) [2](https://www.statista.com/statistics/1340485/usa-most-visited-conservative-websites/), [14](https://virginiatech-my.sharepoint.com/:b:/g/personal/sharonkim213_vt_edu/ERkbQeyerGBGvwPhLMcQ8TkBBPjSOf2qM67ZQbVQKw9oxQ?e=Cc38HT), [15](https://virginiatech-my.sharepoint.com/:b:/g/personal/sharonkim213_vt_edu/EZV2P4IZXlpJpaNvu6LBVIsB0SRd98RdBIeCrX-YxvmmXw?e=oPTdRw)], while for the left group we chose MSNBC and the New Yorker. Given that our focus was to gain familiarity with interaction with the emails, we did not believe that certain selection methods for media outlets would significantly affect our experiment. Our preparation and selection of media outlets can be seen in the excel sheet [here](https://virginiatech-my.sharepoint.com/:x:/g/personal/sharonkim213_vt_edu/EUqh-YVEQ0BCjxsiAmqvOaABrP1Vk89LSLJ9t-ObaEYd2g?e=v7WHdo).

After choosing media outlets, we needed to decide on which newsletters to subscribe to. Our main objective became to sign up for daily newsletters, as well as any current events we thought may be popular (i.e. politics, world news, etc.). In total, we signed up to thirty-two different newsletters from ten media outlets and required ninety-six fake identities ([News\_Question\_Mods - Copy.xlsx](https://virginiatech-my.sharepoint.com/:x:/g/personal/sharonkim213_vt_edu/Ed6bLXvMaBtKswMzhCsnpdYBzlKLA_I0QLFr75KW8z92kQ?e=7WLaLn)). Three fake IDs were signed up to each newsletter to interact in a specific way: (1) do not open emails, (2) open the emails, or (3) open the email and click on a link in the newsletter. All emails were checked and interacted with on Thunderbird, a free and open-source email software that allows access to multiple email accounts. On the days following, we planned to interact and record the data in an excel sheet ([News Coding Sheet.xlsx](https://virginiatech-my.sharepoint.com/:x:/g/personal/elenaroe_vt_edu/ERP9A_8NyaNEr1KzGqwznfcBKkgNccxHZ1gC1--aa11SpQ?e=uwf5VF)) each day except weekends. We distributed all the emails among ourselves to facilitate easy tracking of interactions. The document explaining the experiment and its initial set-up can be found [here](https://virginiatech-my.sharepoint.com/:w:/g/personal/sharonkim213_vt_edu/EcnF5I8gmLFDoD0dKXTsGQsB9NE2xZRkr54mko7RkVAWDQ?e=IrgUa5).

R1: Are there differences in frequency of email received based on the interaction?

R2: What are current barriers to interacting with emails?

Post Sign-Up Summary

For many of the newsletters, only an email address was needed to subscribe to a newsletter and all other personal information (i.e., name, address, birthday, etc.) could be added after creating an account, which was optional. During the sign-up event, we only subscribed to newsletters and did not create accounts. Only a few accounts were created to check the subscription of newsletters. Proceeding after the sign-up event and throughout the interaction stage, we encountered several unexpected results and hinderance. The first being that this experiment was more demanding than anticipated. Even with the expectation of dedicating a few hours to this experiment throughout the week, many scheduling conflicts and prioritization on the birth control paper resulted in half the IDs not being checked regularly. Miscommunication from the technical team also led to small inconsistencies in interaction, which will be further explained in the future works section. Due to this, only IDs 20102-20148 were analyzed in this report with IDs 20134-20148 only being checked starting July 23rd.

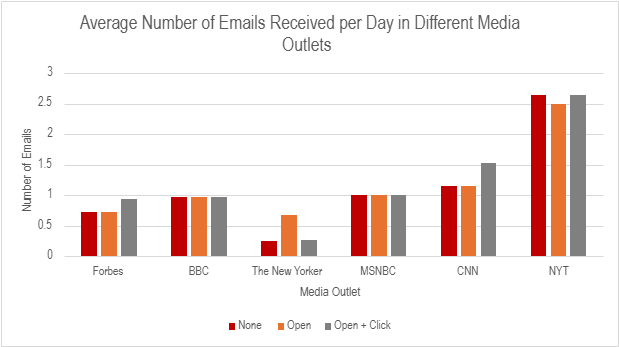
There were also inconsistencies in checking emails daily due to miscommunication from the technical team. All emails on Thunderbird were being “refreshed” or wiped clean after a certain hour each day for a total of three different days. To avoid removing all emails on Thunderbird each day, clear confirmation and communication with the technical team was needed to inform the technical team and have emails moved back onto the software. When emails were being refreshed daily it was challenging to interact with all possible emails received consistently, especially with some emails being sent as early as 3:57AM. If another research question is being done beside this research question again, clear attention is needed to ensure that all emails are being interacted with equally. This may not have significantly impacted our data; however, it made interactions inconsistent throughout the experiment.

Results

*Newsletter Format*

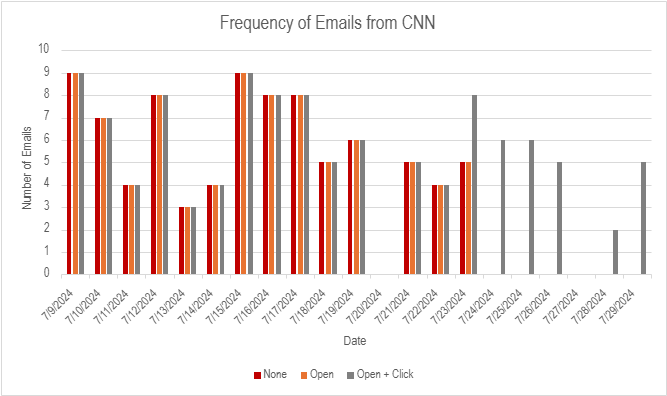
To interact with emails and newsletters, it is important to understand the newsletters and the format. As we understand newsletters more and more, we are then able to better pinpoint how to interact with them. Across all newsletters, except for Breaking News newsletters, the purpose of the newsletters was to highlight key events by adding short summaries and/or links to articles that further explain the event. Generally, the subject line was the title of one of the articles in the newsletters. Some had that article as the first article presented, while others included it in the middle of the newsletter. Breaking news newsletter typically had the title of the article and a link to the article. It typically alerted readers to current issues, therefore for some days there were multiple breaking news sent in a day. Overall, the format of newsletters differed slightly with various articles included.

*Average Number of Emails Received*

Figure 6. Average number of emails received based on each Media Outlet.

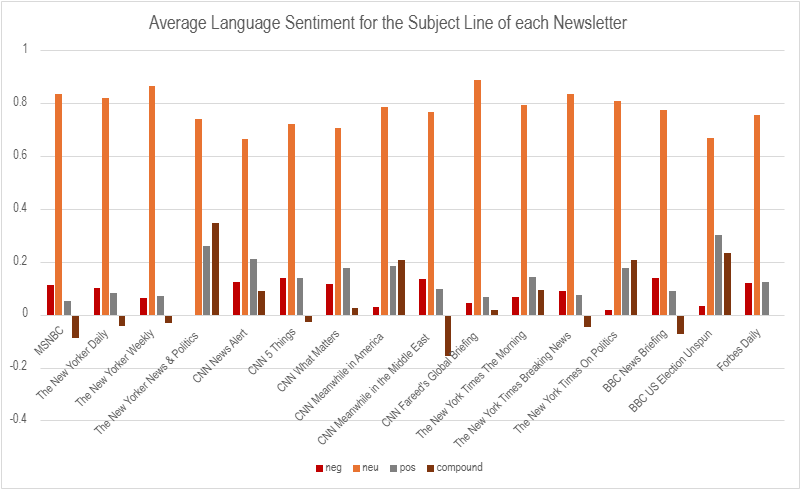
Overall, the New York Times sent the most emails with inconsistent differences within the three interaction groups due to some IDs being signed up for multiple newsletters. Particularly IDs 20137-20139 were all either signed up to the wrong or multiple newsletters. This affected the average number of emails received for the New York Times, despite seeing more emails received from IDs that interacted with the emails. As we went through the fake IDs, we also noticed that some emails were not receiving any newsletters. After further examination into those emails and logging into the accounts, we confirmed that some IDs, such as IDs 20104 and 20106 (the New Yorker Daily newsletters), were signed up for receiving daily newsletters but were not getting any. This was surprising as ID 20105 was receiving daily newsletters and the only interaction done with these emails was opening emails. There were also some that were inconsistent with sending the newsletter. For example, the weekly newsletters sent from The New Yorker did not send their newsletters every week. For IDs 20107 & 20108, one weekly newsletter was received on July 29 (about 3 weeks after the sign-up event). ID 20109, also assigned to receive Weekly newsletters from The New Yorker, received no weekly newsletters.

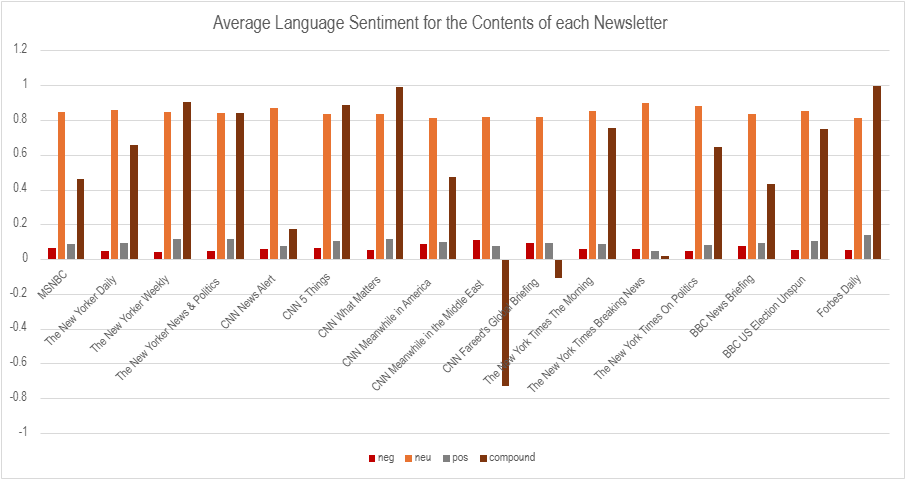
*Frequency*

Figure 5. Total number of emails received each day from CNN.

Based on the interaction with the emails we wanted to see any difference in the frequency of emails we received each day. To our surprise we found slight differences between some media outlets and newsletters. MSNBC and BBC had no difference based on their interactions. Each email address received the same emails with emails only coming from the subscribed newsletter. These IDs did not receive any newsletters from other newsletters or promotional deals. CNN, on the other hand, showed almost no difference between all three groups of fake IDs until July 23rd. After July 23rd we saw that the third group of IDs kept receiving emails from CNN, while the first and second group did not. This was an interesting finding as all newsletters from CNN had the same response. Other than that, almost all other newsletters also sent promotional emails enticing readers to subscribe to them with an ongoing deal and/or sending other newsletters to get them to subscribe. There was a slight notice with IDs that were interacted with (open email and click on an article).

*Language Sentiment*

Figure 3. This chart represents the average language sentiment score for the subject line of all emails that were opened.

Figure 4. This chart represents the average language sentiment score of email contents.

To analyze the language sentiment of each email, an open-source tool, Valence Aware Dictionary and sEntiment Reasoner (VADER), was used. VADER gives four scores: “neg” (negative), “neu” (neutral), “pos” (positive), and compound. The negative, neutral, and positive scores are ratio for proportion, so each score ranges from 0-1 and indicates the proportion of the text made up of either negative, neutral, or positive phrases. The compound score ranges from –1 to 1 (-1 indicating a negative language sentiment and 1 indicating a positive language sentiment) [[11](https://vadersentiment.readthedocs.io/en/latest/pages/about_the_scoring.html)]. Using VADER a python script was created to analyze each subject line and content of the emails received. Because we had to go through each email through Thunderbird, we chose to only analyze the subject line and content email for the ID group that only opened the emails.

Generally, all emails received had a high score on neutral, which can also be seen in figure 3. This may indicate that most newsletters were factual in their subject line and not many were highly biased when presenting their subject line for newsletters to their readers. When comparing the subject line's language sentiment score to the content, the scores remained similar except for the compound score which can be seen in figure 4. The compound score for almost all newsletters is extremely negative or extremely positive. Editors and journalists may want to appeal to a certain group of audiences by portraying some news positively or negatively when it came to communicating through their writings. The specific content may be interesting to analyze as media outlets may focus on different sides of a story, this may be a possible focus for future works. Based on initial observations of newsletters received, there is no change in language sentiment based on interaction. The language sentiment varies across different media outlets; however, it does not vary within a newsletter.

All data analysis can be found [here](https://virginiatech-my.sharepoint.com/:x:/g/personal/sharonkim213_vt_edu/EdX-EiH9DHJAuXBQZuqHtUYBfMW-LQn0B8-m72dNeWBKIg?e=3ZaWnj).

Limitations

The study had several limitations due to various reasoning and factors. First it is important to preface that this experiment is a smaller one that we hope will contribute to the advancement of our current U&A System. More time was spent and needed for the completion of the birth control research question, especially finishing the research paper. With that our proposal for the interactions required a lot of time as we were manually going through each email and recording our data. This acted as a limitation as we were expected to meet deadlines, while balancing this smaller research question. This also confirms the amount of development needed to build an interaction engine. Another limitation we faced was scheduling conflicts among the team that was working on this experiment which prevented everyone from interacting with the emails daily. Not all IDs were interacted with daily which resulted in very inconsistent data, however at least a third of all the fake IDs were interacted with daily.

Future Works

The results and execution of our experiment revealed many opportunities for future work. There are currently many barriers to performing this experiment on a big scale with thousands of IDs signed up for an experiment similar and requiring interaction with. To further the development of our current U&A system and interaction engine for a bigger experiment, we found a few areas for future works. All the areas for future works listed below are areas for building the interaction engine or building a research question studying media outlets.

*Improvement on Interaction*

Time receipts of the newsletters vary with different newsletters. This made it hard to interact with all the newsletters. On top of that, interacting with all the emails takes a lot of time and increases the risk for errors to be made if done manually. The first work with the interaction engine, if not already done, would be to build a system to open all emails daily with either Linux or a python script, more discussion can be done with the technical team. The other area for work is coming up with a way to build an interactive engine that can work similarly to opening an email and clicking on a link. This will require a deeper focus and analysis on understanding what opening an email and clicking an article does and figuring out ways to replicate it. Based on the result, interacting with emails that extends more than opening an email may provide a bigger result if implemented well. In the future, it may also be worth looking at ways to interact with SMS texts and voicemails.

*Defined Characteristics*

Newsletters come in various formats at different times of the day. If we were to do this experiment again, there should be defined definitions of characteristics you are looking for in a newsletter. In our experiment, we gave a generalized implicit definition of what a headline was but led to unclear actions required when analyzing the newsletters. [*Headline*](https://www.merriam-webster.com/dictionary/headline)isdefined as *“*a head of a newspaper story or article usually printed in large type and giving the gist of the story or article that follows” [[16](https://www.merriam-webster.com/dictionary/headlines)]. Generally, the subject line of an email is used to pull in readers. The subject line can be defined as a headline as it informs readers of one story the newsletter will contain. However, with a more specific focus and defined characteristics the study can be done clearly.

*Unsubscribing Experimental Group*

In our experiment, we had three different groups of fake IDs: control group (not opening emails), open emails, & open emails and click on a link. At any point users can unsubscribe to newsletters, however even after interaction with these newsletters do media outlets stop sending these newsletters? Does data continue to be shared?

*Artificial Intelligence*

Artificial algorithms (AIs) have been a highly valuable tool used by many people. The use of AI is even assisting and informing journalists and media outlets to produce articles for their readers. In our experiment we did not analyze in depth the prevalence of AI in the articles and newsletters produced by media outlets. Many media outlets have adopted the use of AI resulting in the development of AI guidelines or policies, with one research study analyzing AI policies from 52 global news organizations [[9](https://osf.io/preprints/socarxiv/c4af9)][[10](https://journalistsresource.org/home/generative-ai-policies-newsrooms/)]. We used [Phrasly.AI](https://phrasly.ai/) to look at multiple media outlets and articles published by the media to see how much AI was detected in each article. However, a deeper analysis of the prevalence of AI is recommended.

*Secondary/Tertiary Rubric*

Initially we decided to look at the most popular/viewed/used media outlets because it seemed the most relevant. However, with platforms like Substack, Revue, and Bulletin, that are open for individual journalists to publish their work, we may study the reliability, trustworthiness, data usage, or personalization of newsletters among sites that are not mainstream. In other words, studying “secondary” and/or “tertiary” sites. More information on creating a rubric for defining secondary and tertiary sites can be found [here](https://virginiatech-my.sharepoint.com/:w:/g/personal/sharonkim213_vt_edu/EYaaeOFgoSBBrNKaXz8FpfEB-YhUWAM3CRrRMiE-xsGQ-Q?e=YPFd7N).

Conclusion

Interaction with emails has shown the potential for research with an interaction engine, however it also showed how much development is needed for the interaction engine. Overall, there are results based on each interaction. Despite our experiment not facing many obstacles and changes, there was still evidence that the interaction does affect the emails received. More specifically, interactions that required emails to be opened and clicking on a link saw the most significant differences in some newsletters.

Research: How can we create an interaction engine that can interact with over 1,000 emails in a way that will drive the media outlets to send an email we want to receive?