1. What is your research question? Explain its significance. Have others looked at this question in the past? If so, what have they found? [We expect 4 paragraphs -- two on the significance of the research question and two that review other's relevant work in this area with citations. The last few sentences should clearly state the contribution you hope to bring to the current state of the literature.]

In the Internet age, rapid information exchange has hugely expedited communication but has also nurtured a breeding ground for the proliferation of misinformation and fake news (Carlson, 2018). As implied in its name, *fake news* is inaccurate information fabricated to resemble conventional news, with the malicious intent to mislead its readers (Mukerji, 2018; Waisbord, 2018). Modern mass media, in particular social media, has significantly lowered barriers, along with providing economical strategies, for third parties to disseminate misleading information (Ng & Taeihagh, 2021). It is estimated that an average American saw and remembered 1.14 fake news months within the 2016 US presidential election (Allcott & Gentzkow, 2017). Moreover, bot accounts that impersonate real humans on social media also exponentially speeded this misinformation by sharing these fake stories (Lazer et al., 2018). These factors combined have enabled fake news to propagate rapidly on social media, posing a danger to the integrity of information exchange on the internet.

Beyond compromising the integrity of communication, fake news also poses significant risks to human behavior. As fake news is commonly created to mislead readers for financial, political, or other gains, it has profound implications (Wang, 2024). For example, during the 2016 U.S. presidential election, fake news exacerbated partisan polarization and conflict (Grinberg et al., 2019). Similarly, misinformation about COVID-19 vaccinations discouraged individuals from accessing essential health resources, disproportionately targeting Americans of color (Lee et al., 2023). Processing fake stories implants false beliefs and propagates misinformation, biasing readers' judgments (Lewandowsky et al., 2012).

In response to these dangers, research has largely focused on evaluating readers' ability to detect fake news (Arin et al., 2023) and developing educational interventions for young internet users (Hämäläinen et al., 2020). However, the field has paid little attention to the intrinsic characteristics of fake news, such as its topics and themes. For example, one study categorized fake news based on its physical properties: physical news content and non-physical news content (An et al., 2023). That is because mainstream research in the field prevailingly assumes that fake news primarily serves as a political weapon, with most of its content rooted in political or *hard news* topics (Allcott & Gentzkow, 2017; Mukerji, 2018; Waisbord, 2018; Van der Linden et al., 2020; Pennycook & Rand, 2021).

However, this assumption may overlook a critical nuance. Mahmoud (2020) found that the novelty of fake news plays a key role in its spread, as novel content captures readers' attention and accelerates dissemination. Importantly, *soft news* (including sensational, entertaining, and cultural topics) tends to be more appealing to casual readers, who are less engaged with political or hard content (Prior, 2003). This suggests that fake news may strategically prioritize soft news content to maximize its reach and propagation, challenging the common notion that it is predominantly focused on political stories. To address this, my research question here is to investigate major content themes in fake news from unreliable sources and compare them with the hard content commonly shared on social media (Bakshy et al., 2015). I answered the following questions to identify similarities and differences between these news types.

1. What are some overlapping and discriminatory words used across documents type?
2. What are major topics (clusters) addressed in both documents type?
3. Does majority of fake news feature hard contents, as assumed in the field?

This investigation holds important implications. If fake news is found to feature hard content predominantly, it would validate prior assumptions. It would also narrow the motivations behind its dissemination, as less engaging content contradicts the goal of a rapid spread. More importantly, identifying recurring themes in fake news could empower readers to approach suspicious stories more critically, especially when these stories fall within identified categories.

2. What data did you choose to address this research question? In what ways does this data shed light on the research question? What might its limitations be? [One paragraph description of the data, several plots or tables of relevant descriptive statistics about your data (e.g. how many documents? what are the length of the dos? tables of metadata? One paragraph discussion of value and limitations relevant to your question.] **Tables are under Appendix**

To address this research question, I used a corpus obtained by a laboratory in the Department of Political Science at UC San Diego, which was collected for an ongoing research project. The corpus consists of 14,543 newsletter emails collected from 181 domains. These domains were categorized into two groups: misinformation websites (106 domains) and hard content websites (75 domains) frequently shared on Facebook (Bakshy et al., 2015). As shown in Tables 1a, 1b and Figure 1, the fake news documents (6,432 emails, 5781 DFM features) were gathered from the misinformation websites, while the hard content documents (8,111 emails, 8,375 DFM features) were originated from the latter category. It is important to note that each domain contributed exclusively to either the fake news or hard content categories. No domain contributed to both the fake news and hard content categories.

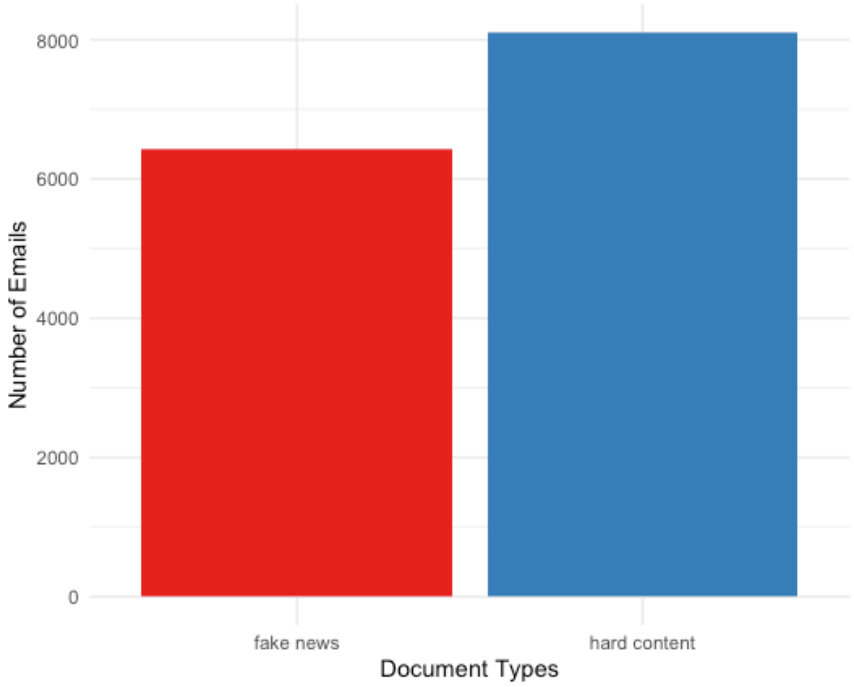

Figure 1. Number of emails from each document types

The emails were collected over a year, from April 2022 to April 2023, with an outage in summer 2022. This collection period was at the latter phase of the COVID-19 pandemic. Presumably, this is a time when online information exchange to be at an all-time high. Therefore, this corpus is highly relevant to the research question and will certainly offer valuable insights. However, there are also has several limitations in this dataset. First, fake news content evolves with world events. Since there was not a presidential election during the collection period, there might have fewer political documents. Instead, anti-vaccination narratives might dominate the fake news content. Second, an outage happened during a period when young internet users, who are typically drawn to soft news content, spend more time online. This could skew the dataset away from such content. Whether these factors counteract each other or exacerbate bias remains uncertain. Finally, newsletter emails are not a universally popular communication method across age groups, it may introduce a potential bias in addressing the research question.

3. How did you go about using the data to create a measure to answer your research question? Why did you choose this approach? [1 page (or however long necessary) step-by-step description of how you went about creating a measure with justification. Why does your particular test answer your research question and make you vulnerable to be proven wrong?]

4. How well did you do at capturing your measure? What were the results of your analysis? [Some metrics and plots that show validation of your measure. Explain your results.].

Preprocessing (Section 3a in codes)

To answer the research questions, I began by preprocessing the data for analysis, ensuring face validity of the research direction. I first separated the dataset into two data frames based on document types, one for fake news documents and one for hard content documents. Afterwar, the body text of documents was converted into a corpus. Preprocessing steps included removing punctuation, numbers, stop words, and applying word stemming to standardize word forms. These transformations ensured that the content of the documents is focused, minimizing noise from irrelevant or redundant features. Two DFMs were created for each document type, capturing the frequency of terms within each corpus. The DFMs were visualized using word clouds to provide a face-valid representation of the content. Figures 2 and 3 depict the word clouds for fake news and hard content corpora. Notably, the word “new” emerged as the most frequent term across both datasets, reaffirming prior findings that novelty is a crucial factor in maximizing the reach and propagation of content (Mahmoud, 2020).

Figure 2. Word Cloud representation for the fake news corpus

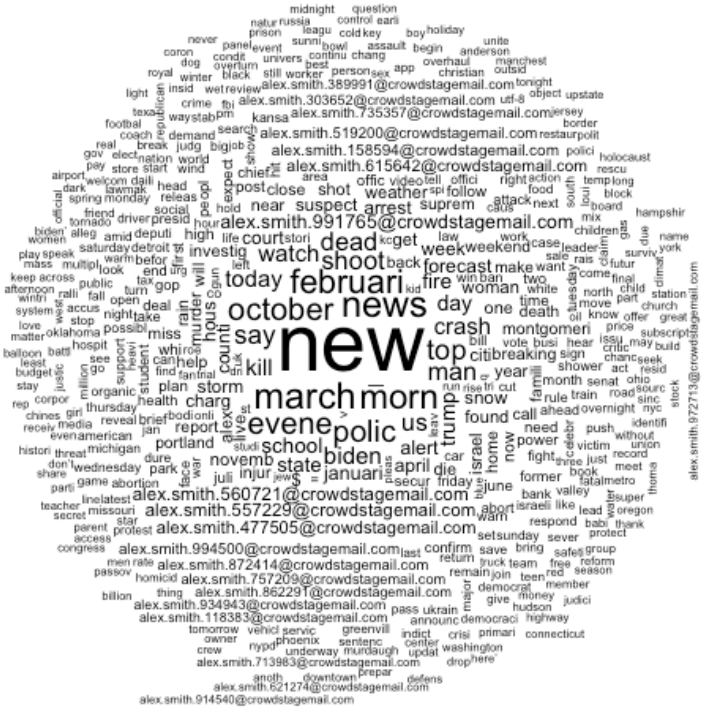

Figure 3. Word Cloud representation for the hard content corpus

1. What are some discriminatory words used across documents type? (Section 3b in codes)

Identifying discriminatory words is a valuable discovery question to start with. This method highlights term that most distinguish fake news from hard contents, offering initial insights into thematic differences, Moreover, dataset already includes predefined categories (document type), which conveniently set stage for discriminatory words analysis. To perform this analysis, the entire dataset was prepressed into a new DFM, which was then trimmed to include only words appearing in at least 20% of the documents. This threshold reduced noise and emphasized more substantively on content features.

Results

According to Figure 4, the top five distinctive words in fake news were: “health”, “ani”, “store”, “product”, and “also”. These terms are subjective in nature and suggest a focus on soft content themes (e.g. sensational health claims or product promotion). This aligns with the hypothesis that fake news prioritizes soft content to attract a broader audience. Conversely, the top five distinctive words in hard content documents are “address,” “book,” “safe,” “add,” and “list.” These terms are more administrative in tone, reflecting attempts to engage readers through incentivizing subscription rather than content. Interestingly, some terms blur boundaries between the document types. For example, “state” (14th most distinctive for fake news) and “world” (18th) seems to align with hard content themes, suggesting that fake news occasionally incorporate stories traditionally associated with hard content. These findings draw ambiguous conclusion, alerting the need for a more complex analysis.

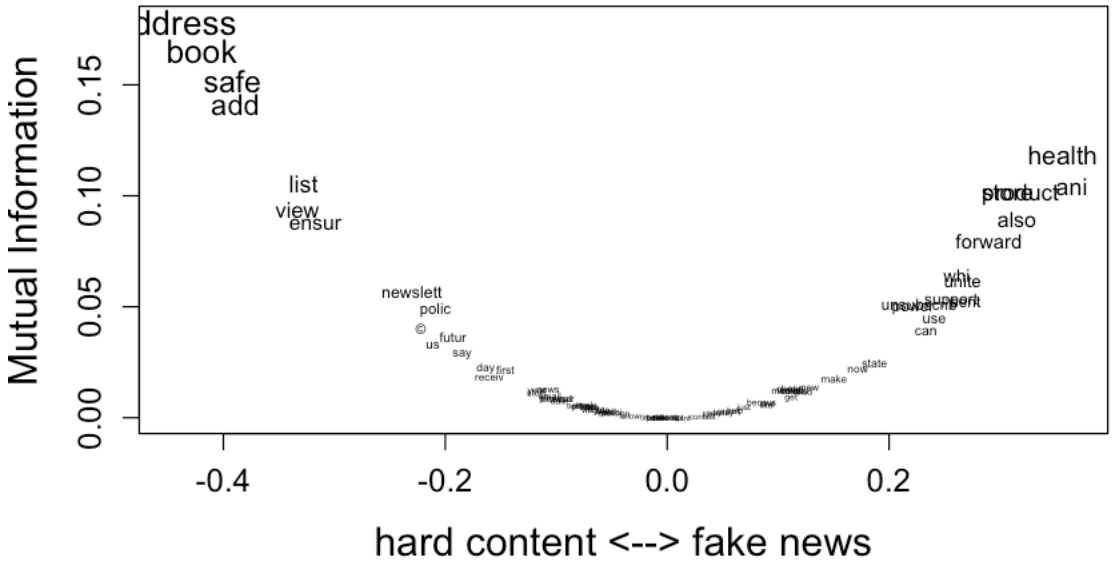

Figure 4. Discriminatory words across document types

2. What are major topics (clusters) addressed in both documents type? (Section 3c in codes)

Topic modeling was employed next to uncover latent themes within the documents. This unsupervised method enables the discovery of cluster without predefined topics, clarifying content similarities and differences. The *K*-means clustering method is an optimal approach in addressing this question. This approach partition documents into *K* topic categories, returning the most probable words for each cluster. The returned most probable words inform underlying themes in these documents.

Using the text processor and prepDocuments function, the corpus for each document type was processed into the Structural Topic Model (STM) format. An initial modeling was first performed with 30 topic categories as demonstrated in class. However, this approach revealed redundancy, as some topics shared overlapping most probable words. Therefore, I gradually

reduced the number of clusters until there are no repeated probable word across topics. The final model included nine topics for fake news and four topics for hard contents.

Results

Table 3 revealed the top five most probable words across nine topics in fake news documents. The probable words for topic 1, 3, 4, and 8 strongly indicated health-related themes, as consistent with the discriminatory word analysis above. Two topics (2 and 5) reflected political themes, with most probable terms like “biden” and “trump”. The remaining topcis (6, 7, and 9) were more ambiguous, featuring structural or stylistic features, unrelated to content.

Fake News	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
Most Probable Word	die	read	health	imag	email	donat	news	organ	post
Second Most Probable Word	der	site	can	email	trump	read	click	health	email
Third Most Probable Word	und	new	cancer	vaccin	biden	minnesota	free	natur	comment
Fourth Most Probable Word	weiterlesen	biden	episod	expos	unsubscrib	news	pleas	pleas	browser
Fifth Most Probable Word	ein	news	watch	covid	materi	today	get	unit	imag

Table 3. The top 5 most probable words across 9 topics in fake news documents

Table 4 revealed the top five most probable words across four topics in hard contents documents. Topics here were narrower in scope, topic 1 reflected international new, while topic 3 and 4 emphasized regional law enforcement stories. Topic 2 features structural or stylistic features, and again ambiguous.

Fake News	Topic 1	Topic 2	Topic 3	Topic 4
Most Probable Word	israel	news	polic	say
Second Most Probable Word	will	email	man	news
Third Most Probable Word	read	newslett	weather	new
Fourth Most Probable Word	can	receiv	counti	polic
Fifth Most Probable Word	now	logo	citi	will

Table 4. The top 5 most probable words across 4 topics in hard content documents

These results suggest that fake news encompasses a broader range of themes compared to hard contents. This reinforces the research motivation that fake news employs diverse content strategies to capture attention, this will in turn benefit propagation. However, both the topic modeling and discriminatory words analysis are frequency-based measures, further validation is required to determine how these words are framed and whether the content align with hard or soft content.

3. Does majority of fake news feature hard contents, as assumed in the field? (Section 4a-4c in codes)

To address the hypothesis that fake news does not frequently include hard content, I sampled 200 documents from the fake news dataset to hand-code if the documents should be classified as "hard content" (1 for yes and 0 for no). The "Hard" category includes documents that factually presented data, studies, or stories. Whether these documents are misleading or is irrelevant to the current classification task. Anything with policy-related commentary were classified as hard contents (i.e., policies, government initiatives, political figures). The geographical coverage of the depicted stories is also irrelevant, in which international news, national news, state news, or county/city news are also included. However, sensational or

entertaining were excluded from this category (i.e., personal anecdote, celebrity-endorsed pseudoscience, advertisement). This manual evaluation focused on the framing and intent of the documents, ensuring that the classification aligned with the research question. The 200 handed-coded documents were then separated into training set and validation sets. Using the training set to train the Lasso regression and Naïve Bayes models, both models were applied to classify documents.

Results

Table 5 reported the precision, recall, and f-score for both model fits. The Naïve Bayes model performed better than the Lasso regression. It achieved a higher F-score of 0.69 compared to 0.52. This indicates a better overall balance between precision (correct positive predictions) and recall (the ability to identify true hard content).

Validation	Lasso	Naïve Bayes
Accuracy	0.74	0.78
Precision	0.78	0.71
Recall	0.39	0.67
F-score	0.52	0.69

Table 5. Accuracy, precision, recall, and f-score for the Lass Regression and Naïve Bayes models

The trained Naïve Bayes model was then applied to the fake news and hard contents datasets separately. Table 6 shows the predicted proportions of hard contents respectively. Predictions showed that around 43.89% of fake news documents were classified as hard content, as compared to 63.58% for hard content documents. These results suggest that fake news do not predominantly hard content, but only occasionally. This aligns with the research motivation that fake news would not have prioritized featuring hard content if their ultimate goal is to maximize reach.

There are also limitations in this classification approach. The training data was sampled exclusively from the fake news documents, which may bias the models' understanding of hard content. This limitation became evident when predicting the hard content documents, as the predicted proportion of hard content (63.58%) was lower than expected. Perhaps one solution is to sample from hard content documents and hand coded for training. However, by doing so may risks overfitting, as majority of the dataset will be classified as hard content, this could lead to high accuracy but low generalizability in the trained models. There are also more available models to train, for instance, using the LLM. However, this approach comes with a prohibitive cost, and was not used in this investigation.

Predicted Proportions	Hard Content	Other Content
Fake News Documents	0.44	0.56
Hard Content Documents	0.64	0.36

Table 5. Accuracy, precision, recall, and f-score for the Lass Regression and Naïve Bayes models

5. How does your analysis shed light on your original research question? What are its limitations and what would be your future work? [2-3 paragraphs.]

The three analyses conducted in this study provide three key insights into the research questions. First, the findings show that fake news documents frequently feature health-related topics, while hard content documents are more focused on administrative goals like subscription prompts. Second, fake news exhibits greater thematic diversity, ranging from political to health-related content, whereas hard content tends to concentrate on international news and law enforcement topics. Finally, the proportion of hard content within fake news was found to be less than 50%, contradicting the prevailing notion that fake news predominantly features hard content. These results highlight the unique strategies used in fake news to attract reader attention, often leveraging soft content themes to maximize dissemination.

This project also introduces two novel perspectives. First, the diversity of fake news content suggests that propagation strategies contradict the emphasis on hard contents. For instance, health product promotions may play a better role in engaging attention. Second, although fake news occasionally incorporates hard content, these stories are not its primary focus. This challenges assumptions that fake news is politically motivated. Instead, fake news creators include diverse content to maximize propagation. Perhaps, dissemination is more a more important goal for fake news creators, than creating political weapons.

I also acknowledge major limitations in the current analysis. The dataset was collected during a non-election period, which may underrepresent politically motivated fake news. This is especially true in the absence of the presidential election, where the heightened partisan polarization was not at play. On the other hand, spam newsletters also likely come with multimedia elements (i.e., videos, images, external links) in newsletters, which was beyond this text-as-data investigation. Future research should address these by examining fake news across different timeframes and analyzing multi-modal content. Furthermore, investigations should also analyze other specific themes for categorization, such as health-product advertisements or celebrity conspiracies. This could explain more about the motivations behind fake news dissemination.

In conclusion, this study provides evidence to reject the assumption that fake news primarily contains politically motivated hard content. Instead, it reveals a nuanced alternative where fake news leverages a mix of soft and hard content to optimize reach, particularly during April 2022 to April 2023.

References

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of economic perspectives*, *31*(2), 211-236.

An, Y., Huang, Y., Danjuma, N. U., Apuke, O. D., & Tunca, E. A. (2023). Why do people spread fake news? Modelling the factors that influence social media users' fake news sharing behaviour. *Information Development*, 02666669231194357.

Arin, K. P., Mazrekaj, D., & Thum, M. (2023). Ability of detecting and willingness to share fake news. *Scientific Reports*, *13*(1), 7298.

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130-1132.

Carlson, M. (2018). Facebook in the news: Social media, journalism, and public responsibility following the 2016 trending topics controversy. *Digital journalism*, *6*(1), 4-20.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 US presidential election. *Science*, *363*(6425), 374-378.

Hämäläinen, E. K., Kiili, C., Marttunen, M., Räikkönen, E., González-Ibáñez, R., & Leppänen, P. H. (2020). Promoting sixth graders' credibility evaluation of Web pages: An intervention study. *Computers in Human Behavior*, *110*, 106372.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094-1096.

Lee, A. Y., Moore, R. C., & Hancock, J. T. (2023). Designing misinformation interventions for all: Perspectives from AAPI, Black, Latino, and Native American community leaders on misinformation educational efforts. *Harvard Kennedy School Misinformation Review*.

Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest*, *13*(3), 106-131.

Mahmoud, H. (2020). A model for the spreading of fake news. *Journal of Applied Probability*, *57*(1), 332-342.

Mukerji, N. (2018). What is fake news?.

Ng, L. H., & Taeihagh, A. (2021). How does fake news spread? Understanding pathways of disinformation spread through APIs. *Policy & Internet*, *13*(4), 560-585.

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in cognitive sciences*, *25*(5), 388-402.

Prior, M. (2003). Any good news in soft news? The impact of soft news preference on political knowledge. *Political communication*, *20*(2), 149-171.

Van der Linden, S., Panagopoulos, C., & Roozenbeek, J. (2020). You are fake news: political bias in perceptions of fake news. *Media, culture & society*, *42*(3), 460-470.

Waisbord, S. (2018). Truth is what happens to news: On journalism, fake news, and post-truth. *Journalism studies*, *19*(13), 1866-1878.

Wang, L. Z., Ma, Y., Gao, R., Guo, B., Zhu, H., Fan, W., ... & Ng, K. C. (2024). Megafake: A theory-driven dataset of fake news generated by large language models. *arXiv preprint arXiv:2408.11871*.

Appendix

Domain names	Number of fake news documents
activistpost.com	19
alphanewsmn.com	626
americanlookout.com	29
americanpolicy.org	6
amtvmedia.com	8
barenakedislam.com	26
beforeitsnews.com	2
biggovernment.news	5
blackeyepolitics.com	141
brighteon.com	134
bugout.news	6
cancer.news	237
carm.org	17
childrenshealthdefense.org	22
christianaction.org	12
clashdaily.com	30
climate.news	7
cloverchronicle.com	3
cnsnews.com	7
collapse.news	9
conspiracydailyupdate.com	424
corona-transition.org	11
dailyexpose.co.uk	133
dailyheadlines.net	1
dailysurge.com	11
davidicke.com	72
dcclothesline.com	9
deplorablekel.com	5
digifection.com	19
djhjmedia.com	1
earthpulse.com	13
ecology.news	6
en-volve.com	10
eugenics.news	6
food.news	4
freedom.news	6
fromthetrenchesworldreport.com	120
gellerreport.com	111
geoengineering.news	181
glitch.news	7

gmo.news	179
gotquestions.org	2
govtslaves.com	9
greatamericandaily.com	136
hannity.com	60
health.news	181
healthnutnews.com	48
hsionline.com	126
ilovemyfreedom.org	38
joebiden.news	10
judicialwatch.org	1
junkscience.com	44
learntherisk.org	12
libtards.news	11
lifespa.com	56
medicalmedium.com	4
medicine.news	181
merica1st.com	1
nationalfile.com	14
nationalrighttolifenews.org	12
nationalsecurity.news	10
naturalmedicine.news	177
naturalnews.com	182
newsinsideout.com	4
newspunch.com	10
newspushed.com	1
newstarget.com	30
nowtheendbegins.com	62
npr.news	12
occupydemocrats.com	53
pandemic.news	10
patriots4truth.org	7
pollution.news	233
powderedwigsociety.com	1
randpaul.news	6
renewamerica.com	4
rt.com	126
savethemales.ca	4
science.news	180
snopes.news	7
steadfastandloyal.com	130
stillnessinthestorm.com	14
swarajyamag.com	197
syrianews.cc	16
tfp.org	6

thecommonsenseshow.com	9
theduran.com	9
thefederalistpapers.org	9
thefreedomtimes.com	14
thegatewaypundit.com	34
thehealthyamerican.org	25
thehighwire.com	150
thehornnews.com	251
thelibertybeacon.com	96
thepublicdiscourse.com	9
thetruthaboutcancer.com	414
thinkamericana.com	2
treason.news	8
trump.news	8
twisted.news	7
unclesamsmisguidedchildren.com	4
vaccines.news	144
vaccinesrevealed.com	68
votefraud.news	5
wakingtimes.com	7
worldhealth.net	56

Table 1a. List of unreliable sources and the number of newsletter emails collected from them

Domain names	**Number of hard content documents**
bangordailynews.com	1
blackamericaweb.com	25
dailysignal.com	505
ecowatch.com	9
electronicintifada.net	5
gawker.com	27
globalgrind.com	4
gothamist.com	17
host.madison.com	10
investigations.peta.org	34
madamenoire.com	4
mediamatters.org	8
nationalreport.net	5
newjersey.news12.com	881
newsone.com	5
secure.nrdconline.org	1
share.credoaction.com	110
sonsoflibertymedia.com	7
talkingpointsmemo.com	30
tellmenow.com	27

therightscoop.com	61
threepercenternation.com	1
unitedwithisrael.org	22
vimeo.com	5
washingtonexaminer.com	92
www.azfamily.com	130
www.barenakedislam.com	24
www.breakingisraelnews.com	400
www.charismanews.com	21
www.cnsnews.com	1
www.commondreams.org	660
www.dailymail.co.uk	107
www.endtime.com	3
www.foxcarolina.com	336
www.foxnews.com	401
www.guns.com	13
www.ibtimes.com	178
www.iflscience.com	4
www.jihadwatch.org	9
www.kctv5.com	663
www.kmov.com	81
www.koco.com	111
www.kptv.com	449
www.lifenews.com	76
www.lifesitenews.com	3
www.mrctv.org	6
www.nationaljournal.com	4
www.nationalrighttolifenews.org	6
www.naturalnews.com	182
www.nowtheendbegins.com	11
www.pbs.org	18
www.politico.com	21
www.politicususa.com	1
www.politifact.com	4
www.rightwingwatch.org	181
www.rollingstone.com	132
www.sfgate.com	26
www.stltoday.com	204
www.theblaze.com	29
www.thedenverchannel.com	1
www.thefederalistpapers.org	5
www.thegatewaypundit.com	28
www.thenation.com	10
www.theverge.com	40
www.timesofisrael.com	283

www.upworthy.com	2
www.washingtonexaminer.com	61
www.washingtontimes.com	53
www.wbaltv.com	112
www.wcvb.com	147
www.wfmz.com	4
www.wfsb.com	114
www.wmur.com	111
www.wsfa.com	307
www.wxyz.com	442

Table 1b. List of hard content sources and the number of newsletter emails collected from them