Understanding Data Privacy Concerns on Reddit Using Unsupervised Topic Modeling and Word Embeddings

Anna Lieb
anna.lieb@wellesley.edu
Wellesley College
Wellesley, Massachusetts, USA

ABSTRACT

In our modern digital age, technology companies and governments increasingly collect and use vast amounts of personal data. Consequently, data privacy has emerged as a salient concern among civilians and public officials. Although multiple existing studies have used traditional surveying methods to assess levels of concern about data privacy, organically-produced web data can provide another dimension of analysis to help us better understand this issue. In order to shed light on recent trends in public opinion and identify specific concerns surrounding data privacy, this study analyzed text data from relevant Reddit posts and comments from a four-month span in 2021. Two methods of analysis were used to interpret this text data: (1) unsupervised topic modeling using a Latent Dirichlet Allocation (LDA) algorithm and (2) a Word2Vec word embedding model trained on the Reddit data. Together, these methods enable us to "discover" specific data privacy concerns that previous surveyors have not addressed. Additionally, by comparing the web data results to traditional surveys, this study suggests new topics of public interest to guide future study and policies about data privacy.

ACM Reference Format:
Anna Lieb. 2023. Understanding Data Privacy Concerns on Reddit Using Unsupervised Topic Modeling and Word Embeddings. In *CS 315 Final Project*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

As personal device ownership and social media usage become increasingly widespread, the presence of digital technology in modern life has continued to grow. This has empowered dominant technology companies and governments to collect, use, and share vast amounts of personal data, with profound implications for our society, politics, and economy. The term "surveillance capitalism" was introduced by social psychologist Shoshana Zuboff to describe the current social-economic system which incentivizes digital platforms' exploitation of mass user data [24]. Even before this term existed, concerns about internet surveillance were present since the beginning of the personal computing boom of the late 1990s and early 2000s [18].

Despite growing public awareness about data surveillance, the collection and use of personal data remains largely unregulated by governments. Some recent privacy laws, such as the General Data Protection Regulation in Europe and the California Privacy Rights Act in the United States, have been introduced to protect consumers' personal data. However, these policies are not comprehensive, and software developers' adherence to these laws is inconsistent [17]. Due to this combination of factors, we now live in a time when about 60 percent of American adults believe it is not possible to go through daily life without their personal data being collected by companies and the government [2]. Assessing the extent and nature of public concern about data privacy is an essential part of strengthening and democratizing the public policy response to internet surveillance [18].

Although multiple surveys have used traditional surveying methods to measure public opinion about data privacy [2, 12], these methods have some limitations. Surveys mostly focus on measuring levels of concern about the general topic of data privacy, without a more nuanced exploration of specific policies and issues within the umbrella of privacy concerns. These surveys are vulnerable to framing effects, which can make assumptions about participants' priorities and exclude different ways of thinking [8].

In this study, we analyzed web data from Reddit forums to build a more complete picture of users' opinions about data privacy. Using Reddit data for this purpose allowed us to learn insights from unprompted, authentic internet conversations between users.[1] This contributes to the larger study of data privacy by mitigating some of the limitations of traditional survey methods, such as framing biases and generality as described above. This analysis could be used to influence the direction of future policy efforts or contribute to existing frameworks for understanding data privacy issues [9]. Additionally, this study reveals salient topics that could inspire future studies and public opinion surveys. We aim to investigate the following research questions:

RQ1: Which data privacy concerns are most commonly discussed by users on Reddit forums?
RQ2: How do these concerns compare to previous public opinion surveys carried out using more traditional surveying methods?

Since this study's approach uses unsupervised learning in the LDA topic model and Word2Vec word embedding, it can "discover" privacy concerns that previous surveyors didn't ask about or address in their survey questions. On the other hand, the data privacy topics that emerge from the results also mirror some aspects of recent survey questions and findings.

2 LITERATURE REVIEW

This section will provide an overview of relevant literature. It includes a definition of data privacy, along with a summary of results

[1] Refer to Section 3.1 to learn more specifically about the features of the Reddit platform and our reasoning for using Reddit data for this study.

CS 315, March 20, 2023, Wellesley College
2023. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

from traditional surveying methods for assessing public opinion about data privacy. Additionally, it outlines previous studies that have used similar methodology for analyzing social media data.

2.1 Defining data privacy

This paper uses the term "data privacy" to describe multiple issues related to the digital collection of personal data by companies or the government. The salience of data privacy issues has surged in recent years due to an increase in the amount of data generated by internet services, social networking sites, healthcare applications, and many other companies [15]. Academic research has identified the twin problems of data privacy and data security as consequences of the emergence of big data [15]. Data security involves the practice of defending information from malicious attacks. Although this is an important issue in our digital age, this paper will focus more on the distinct issue of data privacy. For the purposes of this paper, data privacy can be defined as individuals' rights over how their personal information is collected, shared, and used on digital platforms.

Some legislation governing the collection and use of personal data has also helped to define data privacy in terms of the protection of consumer rights. In Europe, the GDPR requires that users opt in to digital services that collect identifiable user data, among other provisions [13]. Data protection laws are not as robust throughout the U.S.; only some states have data privacy legislation, such as California, Virginia, and Colorado [1, 7, 22]. In contrast to the GDPR's "opt-in" standard, these American laws tend to focus on consumers' ability to opt out of data collection.

Our definition of data privacy captures many of the ideas that are present in this legislation and in academic discussions of data privacy issues. Some research suggests that developers' understanding of the meaning of "privacy" is also influenced by interpretations of policy guidance [14]. Although discourse on Reddit involves a more general population of users who are interested in technology and related issues, news media and public discourse has likely adopted a similar definition of data privacy.

2.2 Traditional data privacy surveys

Multiple existing surveys have done valuable work to establish data privacy as an issue of public concern. One 2019 survey by the Pew Research Center found that a majority of Americans are concerned about the way their data is being used by companies or the government [2]. This Pew survey is extensive and assessed multiple different topics of concern related to data privacy. The four main topics that the study focuses on and the corresponding results are displayed in Table 1. Additionally, the study found that most Americans understand very little about the laws and regulations that are currently in place to protect their data privacy. The survey also found that there were differences in opinion by age group. Young adults were more likely than elderly adults to feel they have control over their data; they were also more likely to think that they benefit from data collection [2]. This is important to note in the context of our study's use of Reddit data, since young males are more likely to be Reddit users [11].

In another study conducted by Gallup in 2018, researchers found that 74 percent of Facebook users were concerned about invasion of privacy while using Facebook [12]. This number had increased

Table 1: Pew Research Center Survey Topics & Results

Percent of Americans with concerns about data collection and use by companies and the government.

Topic of Concern	Companies	Government
Lack of control over collection	81%	84%
Data risks outweigh benefits	81%	66%
Concern over data use	79%	64%
Lack of understanding about use	59%	78%

by 12 percentage points since 2011, which suggests that users' awareness and concern about data privacy has increased in recent years. In contrast to the Pew survey, the Gallup survey did not ask questions about distinct sub-topics relating to online privacy concerns. Instead, the Gallup survey found more generally that participants' anxieties about data privacy had increased in the 7-year period. Additionally, the majority of participants believed that the federal government should pay more attention to internet privacy issues [12].

2.3 Measuring public opinion on social media

Considering the existing assessments of public opinion by traditional survey firms like Gallup and Pew Research [2, 12], one might question the usefulness of a study that uses internet data to measure public opinion. Admittedly, there are multiple limitations to our web data approach. For one, the lack of researcher control over data from social media makes it vulnerable to representation and selection bias [8]. Since our study will use machine learning models, it is also difficult to assess potential sources of algorithmic bias in our analysis. However, using organic web data has unique advantages over traditional surveying methods. Unlike conventional polls, online forums act like open-ended response fields, in which users have the freedom to express their ideas [8]. This allows a new dimension of additional nuance to be extracted from the data.

Previous studies have found that analysis of social media text data can capture large-scale trends in public opinion and correlate with established measures of public sentiment, such as poll data and stock values [6, 20]. We can make use of both web data and survey data by analyzing salient topics on data privacy Reddit forums and comparing these results to conventional survey findings. This provides an ideal balance between the trade-offs associated with each method [3].

This approach draws inspiration from a previous study by Chen & Tomblin [8] that applied a structural topic model (STM) to Reddit data to measure public opinion about autonomous vehicles. In their study, Chen & Tomblin used Pew survey results to interpret the significance of the STM results. In this study, we will use a simpler topic model called the latent Dirichlet allocation model (LDA), in addition to a Word2Vec model. However, we will similarly compare our results to conventional surveys to contextualize our findings.

Although our methodology most closely resembles Chen & Tomblin's study, multiple other researchers have used Reddit text data to study a variety of issues. For example, Yan et al. used a combination of both LDA topic modeling and sentiment analysis on

Reddit comments to compare public sentiment toward COVID-19 vaccines across Canadian cities [23]. Additionally, Li et al. analyzed an Android developer forum on Reddit to investigate how developers take privacy concerns into account in the software development process [17]. In this study, Li et al. used keyword-based filtering and manual refinement methods to isolate relevant Reddit threads for qualitative analysis. This body of previous work provides a foundation for our study's use of Reddit text data and topic modeling methods.

3 DATA & METHODS

3.1 Reddit features and user demographics

Reddit data is ideal for this study for two main reasons. For one, Reddit's features make the platform conducive to text-based dialogue between users. Reddit has a 40,000-character limit on posts, which is high compared to other platforms such as Twitter's 280 character limit. Additionally, unlike Instagram and Facebook, Reddit users tend to use text as the primary mode of discussion. In fact, it is not possible to embed images in the comments of forum submissions, although users sometimes link to images instead. For these reasons, Reddit features encourage users to write more text-based comments and posts, which is important for our document-based topic modeling approach.

Secondly, we chose to use Reddit data because it is more readily available in large quantities than other mainstream social media platforms. Tools like the Reddit API and Pushshift enable researchers to access virtually unlimited numbers of Reddit posts and comments from any time period. In contrast, developer access to platforms like Facebook and Instagram is highly limited. For example, the Facebook API is mostly designed for use by individuals to carry out standard actions on the platform.[2] Twitter's API is more permissive in terms of developer access to data; however, this data is limited to tweets posted within the last 7 days.[3]

One significant challenge that surrounds the use of Reddit data is assessing Reddit user demographics. Since Reddit users are not required to give any personal information when they sign up for an account, it is difficult to estimate the age, gender, race, and other features of users in any given subreddit. However, previous surveys indicate that young, white, and male users tend to be overrepresented in the Reddit community [11]

Additionally, many of the Reddit posts in this study are extracted from domain-specific forums. Two of the most common subreddits represented in the data were r/technology and r/privacy, which are made up of a specific subset of the general Reddit population. Members of r/technology and r/privacy have self-selected to join the subreddit, which indicates that they likely have specialized knowledge and are more opinionated about technology-related privacy issues than the average population. Although this means that our sample is not representative of the general population, it does make it more likely that we will encounter more informed and perceptible opinions.

3.2 Collecting Reddit posts with Pushshift

One obvious data source for this project would be the Reddit API, which is maintained by Reddit. However, a convenient alternative that we will use in this study is the Pushshift dumps, which is a freely-accessible repository of raw Reddit data.[4]

Pushshift is a platform that was created in 2015 to facilitate research using Reddit data. It hosts an archive of all submissions (ie. primary posts) and comments posted on Reddit since June 2005, and continues to collect data daily to add to the archives [4]. One advantage of the Pushshift repositories is that there is no limit on data collection. It also allows researchers to access historical data dating back to Reddit's inception. Pushshift data is organized in data structures that help users easily filter, collect, clean, and store data [4]. The Pushshift dumps offer researchers an extensive amount of acccess to publicly-accessible Reddit data, which will enable us to find relevant posts for analyzing data privacy topics.

In light of previous controversies surrounding researchers' use of social media data [10], it is important to review the ethical considerations of using Pushshift to access Reddit posts at a large scale in this study. Indeed, considering that data privacy is the subject of this paper itself, potential privacy concerns should be taken seriously. The authors of posts collected by Pushshift did not give explicit consent for their posts to be included in this study. However, these posts are publicly accessible to any anonymous user with a Reddit account. Additionally, the data provided by Pushshift is effectively the same as the data that is accessible with the Reddit API, since the Reddit API has relatively relaxed restrictions and is one of the few remaining open APIs among mainstream social media platforms [4]. Therefore, our study's use of Pushshift for collecting Reddit text data is consistent with Reddit's standards for public access.

3.3 Identifying relevant posts

This study will use Pushshift dumps to collect Reddit submissions and comments from a four-month span from May to June 2021. The Pushshift dumps contain posts from all subreddits and all users in this time period, so we will filter out relevant posts based on keywords. To avoid bias in framing effects, we can use a small set of keywords that are most directly relevant to our research questions. Since the inclusion of posts in our data set will depend on these keywords, it is important that we select these keywords carefully.

The term "data privacy" is consistent with our research questions and the vocabulary we have used to define our main research topic so far, so it will be included in our keywords.[5] However, Reddit users may sometimes discuss privacy-related concerns without explicitly using the word privacy. In Europe's primary data privacy legislation, the General Data Protection Regulation (GDPR), "personal data" refers to information that is related to an identifiable person [13]. This term can refer to multiple aspects of data privacy, including data collection, data sharing, data use by companies and governments, and users' control over their own data [17].

For these reasons, this study's data set includes Reddit posts (both submissions and comments) that contain the terms "data" and

[2]Facebook API documentation: https://developers.facebook.com/docs/graph-api/overview/access-levels

[3]Twitter API documentation: https://developer.twitter.com/en/docs/tutorials/getting-historical-tweets-using-the-full-archive-search-endpoint

[4]The Pushshift repository for comments can be found online at files.pushshift.io/reddit/comments/. The repository for submissions can be found at files.pushshift.io/reddit/submissions/.

[5]See Section 2.1 for a definition of data privacy.

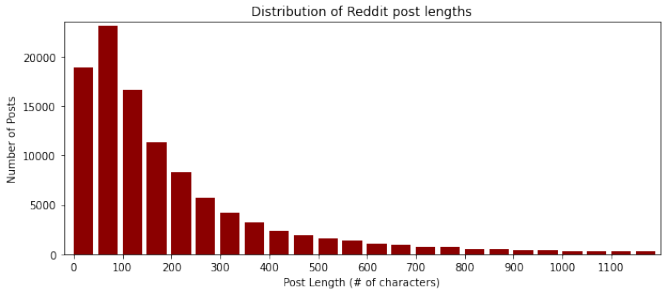

Figure 1: Reddit post lengths

The average post length was 224 words. The max length was 2,137 words, and the min length was 1 word. This distribution includes 106,922 Reddit posts collected from Pushshift repositories. Note that 1,631 posts with 1,200 characters or more are not pictured in this chart.

"privacy" or "personal" and "data." Note that these search terms are not bigrams; instead, the keyword filter includes posts depending on the presence of either pair of words in the text. This keyword selection process was also influenced by qualitative assessments of other combinations of keywords.

Ideally, this list of keywords includes the most common vocabulary in discussions of a variety of data privacy issues, which will enable us to identify the most relevant posts for our research goals. Additionally, our later stages of analysis can move beyond these limited key words. The relevant terms that appear in our initial topic modeling determined the input words to our second round of analysis with word embeddings. Therefore, there is also an element of "snowballing" with key word choice, in which the organic results from one analysis influence future analyses.

Since the Pushshift API contains Reddit data from the start of the platform, we could theoretically collect posts dating back to 2005. However, since this study is exploring current trends in data privacy discussions, we only collected posts from March-June 2021. Future iterations of this study could search a longer time period to incorporate more data and get a larger-scale picture of trends.

3.4 Characterizing the data set

Using the Pushshift repositories, we collected 106,922 unique posts (both comments and submissions) that contained at least one of the keyword pairs "data" and "privacy" or "personal" and "data." The posts ranged in length, but the average post was 224 characters long (see Figure 1 for distribution of post lengths).

A preliminary qualitative exploration of the data indicates that the keywords were overall effective at finding relevant posts. A random sample of these posts is listed in Table 2. Since these posts were posted on different forums across Reddit, a list of top subreddits can offer more insight into where the posts are coming from. A "subreddit" refers to a Reddit forum that is dedicated to the discussion of a specific topic or theme. Overall, the data set contained posts from 10,808 distinct subreddits. The top 8 most common subreddits and their share of posts in the data set are presented in Table 3.

Table 2: Sample of collected Reddit posts

A random sample of Reddit posts containing keywords from March-June 2021. Note that some lengthy randomly-chosen posts were excluded for brevity.

Subreddit	Post text
NoNewNormal	he is not saying that other social medias are better, just the fact that your data is stored in **china**..think about it, one of the most cruel, censoring goverment out there has your personal data (probably knowing what you like better than you yourself) and your response is "other social medias are collecting your data too" oh yeah sure but at very least they respect (to some extend) the data you are giving them and wouldn't just give it out if goverment ask them for it
privacytoolsIO	Just because you're not on the scale of facebook, it doesn't mean that people shouldn't care. That's just deflecting. For a user, being able to delete something is better than being able to delete nothing. The rest of what you say objectively means that you don't delete data at a user's request because it is inconvenient for you. As you said, you're not facebook. In the grand scheme of things, you don't have much data to deal with. Yes, in the end it's up to you whether you allow users to delete everything (while respecting all privacy laws), but in principle what you're doing is laziness at other peoples' expense.
privacy	That would be great. Only if apple wasn't just lying that they care about your privacy. They are just as bad as any other company (maybe worse because of their dishonesty). Only thing they probably do is block everyone else from stealing your data, it's only for apple. People don't care nor know anything about privacy, they just end up believing apple, which is arguably worse than using android and being conscientious about privacy.

4 DATA ANALYSIS AND RESULTS

4.1 LDA topic model results

After collecting the text of relevant Reddit posts and respective comments, this raw text data was processed by tokenizing words, removing stopwords, and stemming words to reduce them to their root. This is standard preparation for the bag-of-words approach involved in LDA topic modeling [5]. The LDA algorithm uses the text data from relevant Reddit posts to create unlabeled groupings of words, which form "topics" based on term frequencies within the text data set [5].

Table 3: Most common subreddits

The set of 106,922 privacy-related Reddit posts contained posts from 10,808 distinct subreddits. This table shows the top 8 most common subreddits and their share of posts in the data set. In total, about 11% of posts come from the top 8 subreddits.

Ranking	Subreddit	% of posts in corpus
1	technology	2.18%
2	privacy	2.15%
3	AskReddit	1.65%
4	CryptoCurrency	1.08%
5	apple	1.07%
6	politics	0.94%
7	news	0.92%
8	worldnews	0.89%

Table 4: Topic interpretations from LDA model

Reddit text data was used to train an LDA topic model, with multiple iterations including both TF and TF-IDF versions. This table includes a sample of interpretable topics and the corresponding term groupings. Topic labels were determined qualitatively.

Topic label	Meaningful terms (selected from top 30 words in the group)	Model type
1. Companies and data	privacy, google, apple, facebook, app, phone, ads, user, companies, sell, security, device, know, tracking, access, account	TF-IDF
2. Governments and data	government, information, law, public, state, privacy, trump, police, china, states, legal, country, new, rights, report, access, national	TF
3. Covid-19 pandemic	vaccine, covid, risk, masks, health, virus, cdc, effects, medical, deaths, cases, pfizer, infection, spread, disease, immunity	TF-IDF
4. Data security online	personal, account, information, comments, email, privacy, breach, post, credit, leaked, message, address, database, phone, online	TF-IDF

Two different iterations of the LDA topic model were trained using the Reddit data. The first was based on a simple term frequency (TF) representation, and the second was based on a term frequency-inverse document frequency (TF-IDF) representation, which weights the term relevance slightly differently. The purpose of running both models was to compare the two to see which would produce the most interpretable categories, and also to increase the chances of identifying multiple distinct topic labels. The results of these topic models are presented in Table 3.

Not every word grouping was included in the table, since multiple word groupings were inconsistent and difficult to interpret. Additionally, some similar word groupings appeared in both the TF and TF-IDF models; for example, both models had categories that roughly corresponded to topics about companies, government, Covid-19, and security concerns as shown in Table 3.

The results of the topic model enable us to broadly identify some of Reddit users' data privacy concerns, which addresses our first research question (RQ1). However, it is somewhat speculative to interpret the specific meaning of these topics without reviewing the original text of the posts. To help address this limitation, we further analyzed the data using Word2Vec word embeddings, which sheds light onto the use of individual words and their meaning in context.

4.2 Further analysis with word embeddings

Topic modeling is effective for identifying high-level trends in the text data; however, labels for these topics rely on interpretation by researchers and the terms within categories are sometimes inconsistent. For these reasons, word embedding models can be used as a second round of analysis to understand our Reddit text data at a deeper level. By introducing word vectorization techniques, word embedding technologies like Word2Vec can help us better interpret the meaning of these terms and the contexts in which they are typically used within the data set.

The Word2Vec algorithm was created by Google researchers in 2013 for natural language processing using a neural network model [19]. This study used Gensim, an open-source library for natural language processing, to model the Reddit post data with Word2Vec.[6] Similarly to the data preparation for the LDA algorithm, Reddit posts were pre-processed for analysis by tokenizing individual words and removing stopwords. This cleaned Reddit text was then input into the Word2Vec model, which produced a word embedding in which words were represented by vectors in a vector space [19]. In the resulting model, words that share similar contexts in the Reddit text corpus are represented by vectors that are close to one another in the space [19].

The results of the word embedding analysis are presented in Table 4. Input words were chosen from the meaningful terms produced in the initial LDA topic model. This methodology is similar to a "snowball" approach, in which terms from the first LDA analysis were further investigated in the Word2Vec model. The output list of words creates an additional set of relevant terms that are used in a similar context to the input word in the text corpus. These output words can be further "snowballed" for future exploration.

The word embedding results indicate that our Word2Vec model worked fairly well, since many of the output words are synonymous to the input word. Since some of the output words are not exactly synonymous, they also offer insights into connections between topics and other words that are used in similar contexts. For example, "companies" as an input word produced direct synonyms like "corporations" and "businesses," but also produced terms that are relevant to the role of companies within conversations on data privacy, such as "advertisers." The word "governments" also appeared on the output list for "companies", which supports the idea

[6]Documentation for the Gensim implementation of Word2Vec can be found at this link: radimrehurek.com/gensim/models/word2vec.html

Table 5: Word embedding exploration

Reddit text data was used to train a Word2Vec model. This table includes a sample of input key terms and the ten most similar words produced by the word vectorization.

Input word	Ten most similar terms
privacy	transparency, security, tos, collection, whatsapp, facebook, apple, anonymity, convenience, decisions
companies	company, corporations, firms, businesses, consumers, customers, industries, industry, governments, advertisers
facebook	fb, whatsapp, instagram, google, advertisers, advertising, tiktok, apple, oculus, ads
sell	selling, sold, harvest, buy, collect, sells, monetize, advertisers, sellers, purchase
government	governments, govt, citizens, ccp, corporations, federal, authorities, agency, foreign, state
law	laws, legislation, courts, legally, legal, gdpr, rules, jurisdiction, regulation, act
gdpr	eu, ccpa, dpa, hipaa, tos, provisions, ico, comply, law, directive

that companies and governments have similar roles in data privacy concerns since they are two major collectors of data.

Further qualitative analysis and interpretation of these results is outlined in the next section (5.1). However, this paper only highlights some of the most notable key takeaways from our results. The results from this study could inspire other insights and interpretations, and interested readers are encouraged to view the code, raw data, and extended results from this study.[7]

5 DISCUSSION

5.1 Interpretation of findings

Here I will attempt to understand the significance of the topics as they relate to existing survey results and data privacy controversies. This helps answer our second research question (RQ2) related to comparing our results to previous public opinion surveys by Pew and Gallup [2, 12]. This can be accomplished by comparing the word categories to topics from survey questions, which reflect how conventional surveys have framed different data privacy issues.

Recent surveys by both Pew and Gallup correctly identified companies and the government as two major actors who collect and use personal data [2, 12]. However, these surveys did not probe further into some of the subtopics that were reflected in the terms in these groupings (see Topic labels 1 and 2 in Table 3). With regard to governments and data, Reddit users mentioned words like "law," "legal," and "rights." The Pew survey investigated some legal aspects of data and found that most Americans understand very little about current data privacy laws [2], but these words in our analysis suggest that laws are still part of the conversation around privacy. Therefore, potential future survey questions about public support of data privacy legislation may prove fruitful.

Both the Pew and Gallup surveys also included questions about data collected for advertising, which is reflected in words like "ads" and "sell" in our topic label 1. However, in questions related to the private use of data, the Pew survey did not ask about specific companies at all, while the Gallup survey asked questions about Facebook, Amazon, and Google. Considering the prevalence of big tech companies in topic label 1, Pew would likely find valuable insights from asking questions about specific big tech companies. Meanwhile, Gallup could update its questions to include Apple among other influential tech companies involved in data privacy issues. The importance of an elite circle of big tech companies is consistent with the "surveillance capitalism" theory for the economic basis of data privacy concerns [24].

Neither Pew nor Gallup mentioned the Covid-19 pandemic in their surveys on data privacy (the Pew survey gathered data before the pandemic in 2019, but the Gallup survey included data collected in 2021). Although the connection between Covid-19 and data privacy is not obvious, the clear emergence of a Covid-19 pandemic category in the LDA topic model indicates that data plays a role in many Reddit conversations about the pandemic. It is possible that this was influenced by the inclusion of loosely relevant or irrelevant posts in the Reddit text data corpus; however, it could also be consistent with heightened news coverage about data related to Covid cases, vaccine efficacy, and tracking of Covid spread by government agencies. The words "cases," "vaccine," "cdc," "risk," and "spread" were all included in this word grouping. Future surveys could do more to investigate the potential relationship between the Covid pandemic and data privacy concerns. Alternatively, instead of asking directly about the pandemic in surveys, researchers could look for relationships between trends in public opinion and pandemic events to see if data privacy concerns may have been changed or reinforced by developments during the pandemic.

When looking to our second round of analysis using Word2Vec, we can find further insights into the topics identified by the LDA model. For example, "China" was included in the initial LDA model, within the topic on governments and data. Then, in the Word2Vec output, similar words to "government" included synonyms in addition to specific mention of the Chinese Communist Party (appearing as "ccp"). Together, these analyses indicate that China is a particularly relevant player in data privacy issues. This could be related to concerns surrounding China's collection and use of data for surveillance practices as an authoritarian regime [21].

A similar second look into government-related keywords can also highlight some salient approaches to government solutions to data privacy issues. "Legal" appeared in the LDA topic for government and data, and the word embedding for "law" produced more related terms including "courts," "legislation," "tos,"[8] and "gdpr." These keywords correspond to some data governance that has been implemented in the past. Both litigation in courts and legislation passed by governing bodies have been used to rein in the power of data collectors [1, 7, 13, 16, 22]. Additionally, requirements for

[7] A repository of these materials can be found at cs.wellesley.edu/ al117/cs315-assignments/finalproject/PM3/ for those with Wellesley permissions. Most of the materials, except for the full Reddit post data, can also be found at https://github.com/annalieb/CS315-data-privacy-reddit.

[8] Note that "tos" appears in the output for "privacy" and "gdpr" as an acronym for "terms of service"

terms of service on digital platforms have been used to encourage transparency for data collection and use. [13]

The EU's data protection law, the GDPR, is one prominent example of data privacy legislation [13]. As one of the only specific pieces of legislation that appeared in the similar words list to "law," it is likely the most well-known example of data privacy law among Reddit users. Using the snowball technique from "law", the subsequent word embedding for "gdpr" outputs other similar policies or legal mechanisms that are leading the way in data regulation and protection. This list could help policymakers identify important policies that are salient in data privacy conversations and existing legislation that could inspire new and more comprehensive laws.

5.2 Limitations

One limitation of this methodology for data collection is that it is difficult to ensure that the extracted Reddit posts are directly relevant to data privacy issues. After qualitatively classifying 100 randomly-selected posts in the data set as either relevant or not, I found that about 53% of the randomly selected posts were directly relevant to data privacy concerns.

One method for improving this relevancy rate could be to apply further filters based on qualities of the posts (such as the subreddit where it was posted, additional key words, or a more complex classification model). I briefly explored the possibility of only including posts from the top 200 most common subreddits in the data set, in an attempt to filter out posts from less relevant subreddits. However, I did not pursue this approach because it limited the sample size to about half of the total corpus (or about 50,000 posts) while only increasing the relevance rate to 62%. A smaller data set could decrease the effectiveness of machine learning algorithms that require lots of text data to be accurate. Further attempts at filtering could also introduce more framing effects, which would be antithetical to the goals of this study.

Since the initial results from the LDA topic model and Word2Vec word embeddings were promising using the whole corpus of extracted data, I continued with my analysis using those results. Our initial results were validated by topics in existing surveys by Pew and Gallup, and also by mainstream understandings of leading data privacy concerns and remedies (for example, the relationship between "companies" and "advertisers" or between "law" and "gdpr"). The scope of this study did not include a more detailed approach to classification of relevant and irrelevant Reddit posts, but future studies could experiment with other, more effective methods for classification using either a rules-based or machine learning approach.

It is also important to mention here that this computational approach for understanding public discourse does not act as a replacement for traditional surveying methods. In a best-case scenario, analyzing text data from the web in this way can offer insights into the topics of conversation, connections between salient issues, or general attitudes related to the issue. However, it is difficult to externally validate such results without further surveying using non-computational techniques and more traditional sampling methods.

6 CONCLUSION

In this study, natural language processing techniques were used to analyze Reddit posts related to data privacy issues. First, the text data was input to an LDA topic model to identify groups of salient terms that formed distinct topics. Next, terms from each of the topics were input to a Word2Vec word embedding model, which had also been trained on the corpus of Reddit data. Together, both the LDA and Word2Vec algorithms helped characterize data privacy discourse on Reddit.

The goal of using these machine learning techniques was to explore data privacy concerns without the framing effects that typically impact traditional question-and-answer survey formats. We found that some of the organically-produced topics mirrored topics that had already been identified by surveys, such as the role of companies and the government in data collection and use. However, we also identified topics that had not been addressed in previous surveys, such as data use during the Covid-19 pandemic, more specific discussions of approaches to data governance, and surveillance by authoritarian governments like China. Our analysis also produced lists of different regulatory approaches and legal frameworks that could be helpful for policymakers or activists who hope to draw inspiration or political will from existing data privacy laws.

Overall, traditional surveying techniques and computational methods should work together to improve our understanding of data privacy and other issues. Since survey questions are often edited to reflect modern trends, computational methods presented in this study could also be used as a preliminary exploration to assess trends in public discourse and guide future survey topics.

Improved iterations of this study could use more robust classification techniques to increase the relevance of Reddit posts in the text corpus. Other future work on this area of study could apply sentiment analysis methods to the text data to assess positive or negative feeling about data privacy actors and issues. A longitudinal study could also contribute to our understanding of how conversation surrounding data privacy has changed over time.

REFERENCES

[1] 2021 Session of the Virginia Senate. 2021. Consumer Data Protection Act. SB-1392. https://lis.virginia.gov/cgi-bin/legp604.exe?211+sum+SB1392

[2] Brooke Auxier, Lee Rainie, Monica Anderson, Andrew Perrin, Madhu Kumar, and Erica Turner. 2019. Americans and Privacy: Concerned, Confused and Feeling Lack of Control Over Their Personal Information. *Pew Research Center* (Nov. 2019). https://www.pewresearch.org/internet/2019/11/15/americans-and-privacy-concerned-confused-and-feeling-lack-of-control-over-their-personal-information/

[3] Eric P. S. Baumer, David Mimno, Shion Guha, Emily Quan, and Geri K. Gay. 2017. Comparing Grounded Theory and Topic Modeling: Extreme Divergence or Unlikely Convergence? *Journal of the Assoc. for Information Science and Tech.* 68, 6 (jun 2017), 1397–1410. https://doi.org/10.1002/asi.23786

[4] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit Dataset. *Proceedings of the International AAAI Conference on Web and Social Media* 14, 1 (May 2020), 830–839. https://ojs.aaai.org/index.php/ICWSM/article/view/7347

[5] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3, Jan (2003), 993–1022.

[6] Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science* 2, 1 (2011), 1–8. https://doi.org/10.1016/j.jocs.2010.12.007

[7] California Consumer Privacy Act of 2018. 2018. SB-1121. https://oag.ca.gov/privacy/ccpa

[8] Kaiping Chen and David Tomblin. 2021. Using data from reddit, public deliberation, and surveys to measure public opinion about autonomous vehicles. *Public*

Opinion Quarterly 85, S1 (2021), 289–322.

[9] Federal Trade Commission. 2021. *FTC Report to Congress on Privacy and Security.* Report to Congress. https://www.ftc.gov/system/files/documents/reports/ftc-report-congress-privacy-security/report_to_congress_on_privacy_and_data_security_2021.pdf

[10] danah boyd. 2016. Untangling research and practice: What Facebook's "emotional contagion" study teaches us. *Research Ethics* 12, 1 (2016), 4–13. https://doi.org/10.1177/1747016115583379 arXiv:https://doi.org/10.1177/1747016115583379

[11] Maeve Duggan and Aaron Smith. 2013. 6% of Online Adults are reddit Users. *Pew Research Center* (July 2013). https://www.pewresearch.org/internet/2013/07/03/6-of-online-adults-are-reddit-users/

[12] Gallup. 2021. *Computers and the Internet.* In Depth: Topics A to Z. https://news.gallup.com/poll/1591/computers-internet.aspx

[13] General Data Protection Regulation. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council. SB-1121. http://web.archive.org/web/20200530095018/https://gdpr-info.eu/art-4-gdpr/

[14] Daniel Greene and Katie Shilton. 2018. Platform privacies: Governance, collaboration, and the different meanings of "privacy" in iOS and Android development. *New Media & Society* 20, 4 (2018), 1640–1657. https://doi.org/10.1177/1461444817702397 arXiv:https://doi.org/10.1177/1461444817702397

[15] Priyank Jain, Manasi Gyanchandani, and Nilay Khare. 2016. Big data privacy: a technological perspective and review. *Journal of Big Data* 3, 1 (Nov. 2016), 25. https://doi.org/10.1186/s40537-016-0059-y

[16] Cynthia J. Larose and Natalie A. Prescott. 2022. Facebook to Pay $90 Million to Settle Data Privacy Lawsuit. *The National Law Review* 12, 129 (2022).

[17] Tianshi Li, Elizabeth Louie, Laura Dabbish, and Jason I. Hong. 2021. How Developers Talk About Personal Data and What It Means for User Privacy: A Case Study of a Developer Forum on Reddit. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3, Article 220 (jan 2021), 28 pages. https://doi.org/10.1145/3432919

[18] Miriam J. Metzger and Sharon Docter. 2003. Public Opinion and Policy Initiatives for Online Privacy Protection. *Journal of Broadcasting & Electronic Media* 47, 3 (2003), 350–374. https://doi.org/10.1207/s15506878jobem4703_3

[19] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[20] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Fourth international AAAI conference on weblogs and social media.*

[21] Alina Polyakova and Chris Meserole. 2019. Exporting digital authoritarianism: The Russian and Chinese models. *Policy Brief, Democracy and Disorder Series (Washington, DC: Brookings)* (2019), 1–22.

[22] The General Assembly of the State of Colorado. 2021. Colorado Privacy Act. SB-21-190. https://leg.colorado.gov/sites/default/files/2021a_190_signed.pdf

[23] Cathy Yan, Melanie Law, Stephanie Nguyen, Janelle Cheung, Jude Kong, et al. 2021. Comparing Public Sentiment Toward COVID-19 Vaccines Across Canadian Cities: Analysis of Comments on Reddit. *Journal of Medical Internet Research* 23, 9 (2021). https://www.jmir.org/2021/9/e32685/

[24] Shoshana Zuboff. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power.* Public Affairs.