Investigating News Source Characterizations using Reddit Audience-based Metrics

Junita Sirait

Submitted in Partial Fulfillment
of the
Prerequisite for Honors
in Computer Science
under the advisement of Eni Mustafaraj

April 2022

Abstract

In this digital age, there is an abundance of news sources, with no one database that characterizes the whole set of news sources. It is a difficult task to provide a meaningful summary regarding the whole set of news producers and to characterize each of them, due to their massive number and the domain knowledge needed to manually characterize them. In this thesis, I investigate the task of news sources characterization using only their Reddit audience sharing practices and metrics as the features of the news sources. In total, I include in my analysis 2,647 news sources represented by 2,189 subreddits, with sharing metrics including the number of Reddit submissions that mention those news sources, the number of associated comments, and the total upvote ratio of those submissions.

By visualizing the representations of these news sources built using their Reddit audience-based metrics, I find that the news sources that are close together in the representational space indeed have some "similar" characteristics, although the notion of similar is different from one neighborhood of news sources to another. Therefore, for unlabeled news sources, one can potentially observe where they are located in the representational space built with the Reddit audience-based metrics as outlined in this thesis, and infer particular characteristics of those unlabeled news sources by leveraging known facts about the neighboring news sources.

In addition to the visual examination of the representations of the news sources, I also attempt to cluster them together using their Reddit audience-based sharing statistics. Although imperfect, the clustering results suggest that particular news source characteristics such as country of origin and some specific themes are prominent features that give rise to cluster structures that are easier to identify using k-means algorithms, than other features. Similarly, a small case study to build a simple classifier suggests that building a classifier for predictive tasks is easier accomplished for some features than others.

I find that there are some evidence that Reddit audience sharing statistics alone can be used for inferring some characteristics of news sources. Even a simple exploration of comparing the total sharing frequency of a news source and the number of subreddits mentioning those news sources leads to interesting insights about news source popularity and the broadness of their audience, as shown in the Exploring News Sources section. However, more work is needed to further engineer or improve these features and potentially build models to predict these characteristics for unlabelled news sources.

Acknowledgments

I am incredibly grateful to Professor Eni Mustafaraj, who has not only advised my honor thesis, but also given unwavering guidance, support, and encouragement throughout my undergraduate journey. What a good fortune it is to have Eni as a professor, advisor, and mentor.

I would also like to express my gratitude to Professor Cassandra Pattanayak and Professor Sohie Lee for being on my thesis committee and for their insightful advice and feedback. Thank you to Professor Ann Trenk for graciously agreeing to be the Honors Visitor on my thesis committee.

Thank you to Wellesley Cred Lab members, especially Beatriz Paulino, Ashley Jang, Annabel Uhlman, Veronica Lin, and Ropah Shava, who have helped me greatly in my data collection and labelling processes.

To my friends who have patiently listened to my rants and ideas, and been there to offer encouragement and support – thank you!

My most tremendous gratitude to my Dad, Mom, and Brother, for providing limitless support and warm comfort throughout my life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Engaging with news is how we learn about events and information originating in various places all around the globe. In this digital age, news is abundant as the number of news producers, aggregators, and spreaders is multiplying thanks to the advancing digital technology, which lowers the technical barrier of participating in the online news ecosystem.

However, even with the same advanced technology, there is not an organization or database in the world that is thoroughly keeping track of these prominent actors in the online news ecosystem. It is a difficult task to provide a meaningful summary regarding the whole set of news producers and to characterize each of them, due to their massive number and the distributed nature of how the Web works.

Manual investigation to summarize and characterize each of them is simply too expensive in terms of time and labor. Automated ways of doing so have been attempted (largely by companies or long-term big research projects), by characterizing not only based on the self-reported labels of news sources, but also the actual news articles that are produced by them . Yet this is also resource-expensive in terms of the computing power that is needed to do so, and the lack of high quality training data. As a consequence, the resulting database of news source characterizations is either only available as a paid service, incomplete, or hard to replicate.

Generally, the efforts of news sources characterizations can be divided into three main categories: a) content-based, b) audience-based, and c) propagation-based.

The efforts that I listed in the previous paragraph are mainly concerned with content-based characterization. They depend on Natural Language Processing algorithms, which often need huge computational power and big high-quality training datasets. My thesis, on the other hand, aims to investigate audience-based characterization of news sources. Here, the term "audience" is defined as people who engage in reading, sharing, commenting, liking, and discussing news on the internet. These engagements take place in various online platforms, for example Facebook, Twitter, or Reddit. This thesis focuses on the Reddit audience, who discuss and engage with news articles in theme-specific communities on Reddit, called "subreddits".

By focusing on the audience of the global news sources, I aim to investigate whether we can use this audience-based approach in answering questions about the news ecosystem in a large view or individual contexts. For example, we could investigate questions from the following list:

- What news sources are most popular across different subreddits?
- Which ones are shared the most frequently regardless of the subreddits (i.e. largest sharing volume)?
- Can we infer the reliability of news sources based on their co-occurrences with other news sources in various subreddits? What about their political biases?
- What is their primary medium (e.g., TV, radio, Web)?
- What is their country of origin and language?

Combining these kinds of questions, I formulate this research question:

RQ: *Can we use news source audience engagement on Reddit to create meaningful representations, which in turn will allow us to investigate a series of characteristics of news sources using unsupervised and supervised techniques?*

This thesis, is my attempt to answer this question. In this document, I first go over some background information about news sources, news sources characterizations, and Reddit in Section 2. Then, in Section 3, I connect my work to existing literature on

the topic. In Section 4, I introduce the source of my dataset to build a dataset of sites I recognize as news sources, as well as audience engagement metrics for their representations. In Section 5.1 I investigate the total sharing frequency of news sources and compare it to the number of subreddits in which those news sources are shared, to characterize the popularity and broadness of appeal of those news sources.

In Section 6.2.2, I build news sources representations using the subreddits that discuss them as their features, representing them as a matrix. I then explore and process this matrix including by filtering subreddits in Section 6.2.3, filtering news sources in Section 6.2.4, and employing the dimensionality reduction algorithm PCA in Section 6.2.5.

In Section 7.1, using the vector representations of the news sources, I employ t-SNE to visualize news sources in a three dimensional space and I visually investigate whether news sources that are closer together (have short distance between them) share similar characteristics. In Section 7.2, I also attempt to cluster the news sources based on their vector representations, to evaluate if the news sources that are close together have any similar characteristics. If they do, then we can reason that embedding news sources based on their audience, using the method outlined in this thesis, does indeed yield in news source representations that place news sources with certain similarities together. Knowing this, for the unlabeled news sources that are in our dataset, we can infer particular characteristics for them, based on their close neighbors in the vector space. As a small case study, in Section 7.3 I attempt to build classifier models to predict some characteristics of news sources, particularly their country of origin, and their reliability.

In Section 8.1 I summarize my findings, and I outline future works to undertake in Section 8.2.

Contributions of this thesis are as follows:

- In building a set of news sources, I find that there is little overlap between what are considered as news sources by the freely accessible GDELT project, and the company Muck Rack, confirming the need of a deeper understanding and more complete record of all available news sources.

- I investigate a novel way of news sources representation by solely using their sharing statistics on Reddit, and explore its usefulness in news source characterization tasks.

- I provide evidence that interactive visualizations such as t-SNE are useful tools for answering questions about the similarity of news sources.

Chapter 2

Background

In this chapter, I discuss how this thesis defines "news" and "news sources", I describe the task of news source characterization, and I introduce the social platform Reddit and the data that I acquire from it.

2.1 About News and Who Produces It

According to Wikipedia, "News" is the reporting of current events usually by local, regional or mass media in the form of newspapers, television and radio programs, or sites on the World Wide Web [1]. The news ecosystem is formed by complex interactions of actors in it. For example, news producers produce news stories. These news producers are what I refer to as "news source" or "news media". Some of these producers are independent, for example BBC, CNN, NYT, etc., but some are hosted on widely-accessible platforms such as Facebook, Blogspot, Apple Podcasts, WordPress, etc. The news stories that they produce are aggregated by yet another type of agents, such as Google News (which relies heavily on algorithms), HuffPost (which relies on journalists), and others. Finally news is consumed and shared by news audiences, such as Reddit users. This thesis is concerned with characterizing news sources, which are the actual actors that produce and publish news with their own editorial boards, and will disregard news aggregators and hosts.

[1] https://en.wikipedia.org/wiki/Category:News

It is also important to mention that currently there is an abundance of user-generated content, such as blog posts and podcasts, that discuss and share diverse materials including news. These user-generated contents are commonly hosted by sites such as Medium, Facebook, Blogspot, Apple Podcasts, WordPress, etc. In this thesis I will disregard both as they are not inherently news sources with an independent board of editors and journalists.

In this digital era, online news reporting agents are abundant, and different institutions have different definitions of what counts as a news source. Instead of trying to come up with a novel definition of news source, in my thesis I rely upon the definition of news sources used by third-party organizations, in particular Global Database of Events, Language and Tone (GDELT) Project [2], and Muck Rack [3]. More specifically, I only include sites in my set of news sources if they exist in both the GDELT and Muck Rack databases. I introduce both GDELT and Muck Rack in Section 4.1.1 and Section 4.1.2 respectively. After acquiring the intersection of the set of news outlets recognized by GDELT and Muck Rack, I then filter out some major news aggregators, user-generated contents, and hosts, for reasons described previously. I further explain the reasoning and process of filtering news sources in the Section 5.1.

2.2 News Source Characterization

News source characterization tasks include characterization by media type (e.g. print, online, podcast, video, etc.), coverage reach (e.g., local, regional, etc.), topic (e.g., politics, sports, etc.), geographical origin, language, reliability, political leaning, and others. Some of these characterizations are easier to do than others, for example characterizations by media type and language, as one can infer these characteristics explicitly as they consume news from any news source. Characterization by geographical origin is also easy for the news sources that self-report their geographical origins, which most if not all of them do. Characterizations by reliability and political lean-

[2]https://www.gdeltproject.org/
[3]https://muckrack.com/

ing are harder to do, as one has to have domain-specific knowledge to characterize manually, or a model has to have gold-standard training data and be fairly complex to characterize automatically. Non-manual characterizations have mostly been done by looking at the style of writing [18] or the content of the news itself using Natural Language Processing (NLP) [12], or the features of the news sources themselves such as their social media accounts, URL structure, and types of Web traffic they attract [2].

In this thesis, I ponder whether we can infer these characteristics of news sources using audience-based metrics. Yet another question to answer is: based only on their audience, how are news sources characterized as "similar"? Which characteristics are easier to infer based on audience behaviors, and which are harder to infer? These are the questions that this thesis attempts to answer.

To characterize news sources, in this thesis I attempt to build vector representations (embeddings) of my set of news sources using audience-based data that I construct from Reddit posts. I then investigate whether various characteristics of news sources can be inferred using their vector representations by clustering them, and building regression and classifications models to predict their characteristics.

2.3 Reddit

When considering audience-based metrics, researchers often turn to social media platforms such as Reddit, Facebook, and Twitter, where news is shared abundantly. This thesis specifically considers Reddit and this section serves an overview of what Reddit is as a platform and why it is chosen here as a data source to build audience-based metrics to characterize news sources.

2.3.1 Reddit as a Platform

According to Wikipedia, Reddit is an American social news aggregation, web content rating, and discussion website [4]. It was founded in 2005 and, according to Alexa, is

[4]https://en.wikipedia.org/wiki/Reddit

currently one of the most used social platforms, with 40.5% of its users residing in the USA [5]. Reddit also ranks at the 23rd highest traffic globally, and 6th nationally in the US; higher than Twitter (ranking 27th and 18th respectively), another social platform that people commonly use to discuss news among others. As of March 2022 Reddit has more than 50 million users [6].

One starkly unique feature of Reddit is that it provides the opportunity for users to be anonymous. Unlike Facebook that requires users to use their real names and identities as well as create only one account[7], Reddit allows users to preserve their anonymity by not asking personal identifiers[8] (except email for signing up[9]), and creating more than one account (as long as they are not used to mass-vote a post)[10].

Reddit is accessible by everyone, even without an account. However, one needs an account to interact with posts and other users on the platform, such as to post, comment, vote, share, give awards, or send a private message. A post on Reddit is also called a "submission", and I will use these terms interchangeably. Users generally post to "subreddits" which are dedicated forums or communities on Reddit that are topic specific, although users can post on their own profile page, if they so wish. They can comment and vote on submissions, as well as reply to comments and vote on the comments themselves. Users can also give "awards" to posts or comments as a sign of high appreciation, however users need to pay to access and give those awards. Users can also share posts as well as report them if deemed inappropriate.

Subreddits themselves are created by a Reddit user and moderated by one or more Reddit users. An example of how a subreddit looks like is shown in Figure 2-1. In addition to regular posts by Reddit users, subreddits can also contain advertisements, which are a source of revenue for Reddit. Advertisements are not always related to the theme of the subreddits, but users can interact with ads the same way they interact with posts: upvote, downvote, comment, give awards, or share.

[5]https://www.alexa.com/siteinfo/reddit.com
[6]https://www.redditinc.com/
[7]https://www.facebook.com/terms.php
[8]https://www.redditinc.com/policies/privacy-policy
[9]https://reddithelp.com/hc/en-us/articles/360060420092
[10]https://reddithelp.com/hc/en-us/articles/204535759-Is-it-ok-to-create-multiple-accounts-

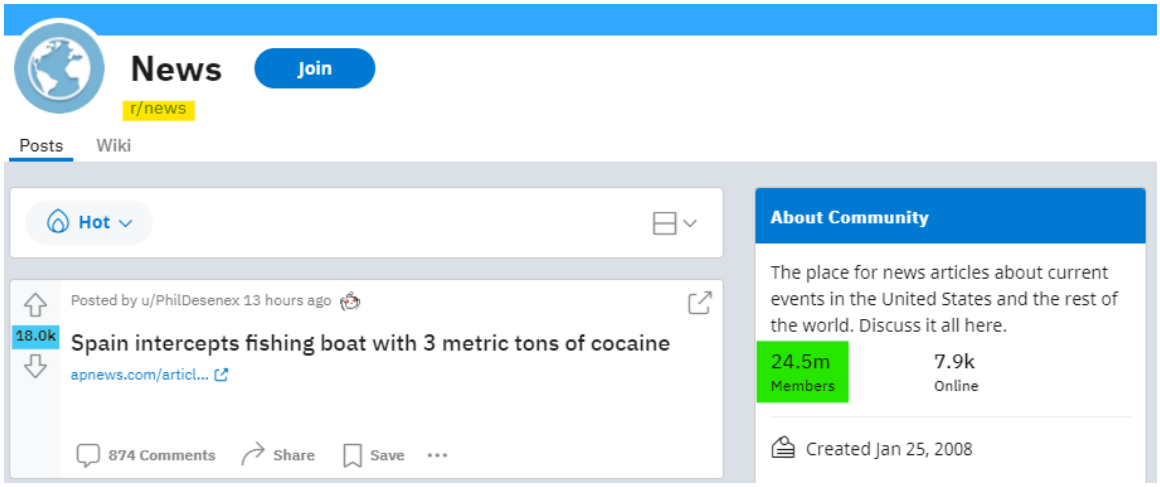

Figure 2-1: The page of subreddit r/news (as highlighted in yellow). It contains submission space, as well as details about the subreddit such as descriptions, number of members/subscribers as highlighted in green as well as subreddit-specific rules (not shown here). The first post that shows up has 18k total votes (as highlighted in blue), which are the highest among the most recent posts.

Each subreddit has their own rules, such as for example discouraging spams and encouraging tagging posts with a descriptive word (e.g. [Advice], [Looking for], [Available Lease] tags in r/nycapartments) in addition to the general guidelines of respecting others. Moderators of these subreddits are users (or bots) that make sure that all posts on their subreddits adhere to their subreddit rules. Moderators have the right to de-list posts that are deemed inappropriate.

Users can join as many subreddits as they want. This thesis will use the term "join" and "subscribe" interchangeably. The official term used to be "subscribe", but Reddit changed it to "join" a couple of years ago[11]. However, the database from which I gather my data (Pushshift) still uses the term "subscribe". The Pushshift dataset will be introduced in Section 4.2. When a user subscribes to a subreddit, the submissions on that subreddits will appear on the user's Reddit timeline.

There are four types of subreddits: public, restricted, private, and premium-only[12]. For public subreddits, users can both vote and comment on a submission, regardless of whether the user is a subscriber of the subreddit in which the submission is posted

[11]https://www.reddit.com/r/modnews/comments/b698ao/

[12]https://www.reddithelp.com/hc/en-us/articles/360060416112-What-are-public-restricted-private-and-premium-only-communities-

or not. In restricted subreddits, anyone can view and comment but only approved users can post. These users are approved by the moderators. Private subreddits are only accessible to approved users, so no other users can view and participate in these communities. On the other hand, only Reddit Premium members can create, view, and participate in Premium-only communities.

There are two voting options on Reddit: upvote and downvote. Upvote will add +1 point to a post while downvote adds -1 point. A high number of upvotes will give the associated post more visibility as the post is ranked higher and will appear higher on the subreddit page (as seen on Figure X above) and the timeline of the subreddit subscribers. In general, posts are ranked by both the total votes they have as well as when they are posted. When a user makes a post, Reddit automatically sets the total vote of the post be +1, with the user themself being the first voter.

In addition to voting on posts, users can also vote on the comments associated with those posts. The technical details are the same, in that upvote will add +1 point to a comment, while downvote adds -1 point. Comments can be sorted based on various metrics, including based on the total votes they have. Although comments are an interesting part of Reddit to study, this thesis will focus more heavily on Reddit submissions that contain the links of our set of news sources. This thesis takes into account the number of comments associated with each submission.

Although users are free to post submissions, Reddit also has a system in place to regulate users" activity, called "karma", which is a kind of score a user receives based on the total votes that their posts and comments received. For example, if a user posts in a subreddit and receives 5 upvotes and 1 downvotes, they would gain 4 karma points. Higher karma points allows users to post more frequently. This also means that new users have limited posting ability, at least until they gain karma points. An example of user profile is shown in Figure 2-2.

In addition to the regular free Reddit account type, Reddit also offers paid premium account type for users. For $49.99 per year, users can upgrade their accounts to premium, allowing them to access Reddit without ads, as well as to use extra features such as filtering posts by subreddits or topics or pinning posts on their homepage.

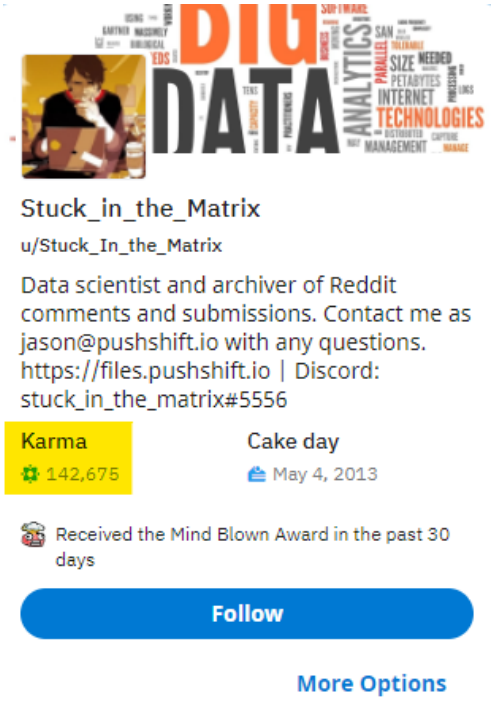

Figure 2-2: An example of a Reddit user's profile. u/Stuck_In_The_Matrix is the Reddit username of Jason Baumgartner, who is the creator of the Pushshift dataset. The amount of Karma he has is highlighted in yellow.

In general, the members of a subreddit are very dedicated to the cause or theme of their particular subreddit. Such dedicated participation has been publicly shown, for example, by the event of the short squeeze of various stocks, most notably Gamestop (GME), initiated by small-scale investors notably organized in r/WallStreetBets, which ended up affecting the financial system significantly [7]. This shows how interactions and discussions in subreddit not only reflect the real-world, but also affect it. For example, Zannettou et al. revealed that some alt-right communities in Reddit could have significant influence in spreading alternative news to the other social platforms such as Twitter and the Web in general, thus affecting the larger public communities [45]. This confirms the important link between Reddit and our society and thus supports the utilization of Reddit as a tool to gain insights about various aspects of our society.

2.3.2 Reddit as a Data Source

Reddit data is valuable because it is the most permissive social platform to access. Meanwhile, other social media platforms like Facebook and Twitter, unlike Reddit, have gotten less permissive over the years as companies juggle their priority of users" data safety and commitment to openness and transparency. Since Reddit users are largely anonymous, Reddit does not have as many privacy concerns which leads to a more accessible source of data.

Anonymity itself is an important feature to consider as I aim to answer my research question of characterizing news sources based on their online audience. According to Luarn and Hsieh, users are more willing to express opinions in anonymous conditions than in non-anonymous conditions [27]. This means that the opinions that people express on Reddit are more likely to be more representative of how people really feel. As such, the news source consumptions and sharing patterns are reflective of what people's preferences are in the physical world.

Additionally, Reddit also provides arguably familiar forums where people discuss common interests, in the community-driven subreddits. According to Luarn and Hsieh, this also encourages "communicative discussions" and "the expression of unconventional views" which further confirm that people's true preferences towards news sources are reasonably correctly mirrored on Reddit.

The abundant data and users, great accessibility, and built-in anonymity features of Reddit makes it a really attractive platform to study various issues in our society. For example, Kumar et al. utilizes Reddit data to quantitatively examine society's early perceptions of the COVID-19 vaccines by using topic modeling algorithms to understand how dominating words related to COVID-19 and vaccinations changed overtime [23]. Choudhury and De specifically use Reddit data to examine the role of anonymity in discussions about stigmatic mental illnesses and the general role of social web in behavioral therapy [10].

Moving from the medical field, researchers also utilize Reddit data to study social issues and phenomenons. Using Reddit data, Fire and Guestrin examine how trends

and network stars emerge and fade overtime by creating large temporal networks and investigating how high-degree vertices emerge and dissipate, which could for example give insights to how people gain and lose political power, and how disease spreads throughout populations [16]. Veselovsky et al, on the other hand, utilized Reddit data to develop neural embedding methods to understand the social context of music sharing such as how there are a lot of extra-musical factors in music sharing [43]. Yet another use case of Reddit data is demonstrated by Setty and Rekve, where they utilized Reddit data for analysis and detection of fake news, using both the textual content of and social media comment to detect false claims [38].

There are many other published research projects that rely on Reddit to gain insights about various aspects of our society, justifying the use of Reddit data in these types of social computing study. Furthermore, as news plays an increasingly important part in our society, I see it fitting to utilize Reddit data as a source of social behavior in news consumption, to study news sources.

Chapter 3

Related Works

News sources do not generally come with identifying labels such as political leanings/biases, credibility, audience targets, and types/genres of news stories in focus, thus it is impossible to fully understand their characteristics if we only rely on their self-reported characteristics. Instead, the efforts of news sources characterizations can be divided into three main categories, based on content, audience, and propagation. In my thesis, I will focus mostly on the intersection of the second and third approach for news sources characterizations. This is because in this thesis I look at how news sources audiences share or propagate news links in various subreddits.

Online communities in general, albeit being mostly anonymous, prove to be a useful indicator/feature to consider, in the effort of understanding the online news ecosystem. As shown by Wang et al, "different communities discuss different types of news" and that some fringe communities could have "disproportionate influence with respect to pushing narratives around certain news" [44]. In this thesis, in addition to gaining insights about news sources based on the subreddits they are shared in, I also explore the diverse subreddits themselves.

Past research projects have utilized the sharing patterns of news stories" links on social media platform to infer various characteristics of new sources, such as their political leanings (by Robertson et al.[34] and Bakshy et al.[1]), reliability (by Pennycook et al.[32]), and their social genre (audience-based clusters) (by Samory et al.[37]). I will go through the overview of each research project in order to summarize and

compare the findings and methodologies used so far in this effort of gaining insights about the online news ecosystem.

In their study, Robertson et al. investigated the partisan bias (political leanings) of various components of Google's Search Engine Result Pages (SERPs) since people extensively use Google search engine to find and fact-check information [14, 13] and that partisan bias could affect undecided voters" decision in the 2016 election [14]. However, it is hard to create an objective partisan bias score that can be used for their purposes of auditing the search engine (Google) with respect to the existence of a "filter bubble" phenomenon. Establishing an objective partisan bias score is hard because, if done manually, this task is expensive in terms of time and labor and risks subjectivity. In the paper, they tackle this issue by deriving audience-based partisan bias scores for web domains, specifically by leveraging the social network principle of homophily observable in Twitter, Facebook [1], and thus arguably other social platforms such as Reddit. To derive the partisan bias scores, Robertson et al. uses the voter registration data to label people's political leaning, find them on Twitter, and use their tweets to investigate the URLs they have shared. The actual Partisan Audience Bias (PAB) of a news source domain is calculated by the scaled difference of the proportion of democratic accounts and republican accounts sharing the links from that domain, resulting in partisan bias score in the range of -1 (shared exclusively by democrats) and +1 (shared exclusively by republicans), where PAB score of 0 means that the links to that particular news source were shared by the same proportions of democratic accounts and republican accounts.

Robertson et al.'s PAB scores have a $r>0.9$ correlation with Bakshy et al., and similarly positive correlations with other partisan-bias-scoring projects such as All-Sides ($r>0.6$) and Pew Research ($r>0.7$) [34]. This shows that audience-based (thus community-based) characterization of news sources are comparable to content-based and rater-based characterizations [34], justifying the use of Reddit's communities (i.e. subreddits) to empirically characterize news sources. Furthermore, the use of the largely anonymous and community-driven platform, Reddit, alleviates the issue of competing power between audience's choice in consuming/sharing news stories and

the platform's algorithm effects.

In addition to political leaning (partisan bias), the audience role has also been utilized to infer news sources" reliability, as shown by Pennycook et al. [32]. In their paper, Pennycook et al. use survey-based methodologies to show that the average person's judgment of the reliability of news sources correlates strongly with expert-based reliability-check for the domain-level judgment (as opposed to story-level). This is shown to be true (almost) regardless of the political leanings of the audience. Although the research methodologies do not rely on audience sharing patterns, this research still showcases the importance and strength of the crowd, suggesting that it is beneficial to infer insights from the news stories audience in general, arguably including their sharing patterns.

Characterizing news sources based on political leanings (partisan bias score) and reliability provides useful insights to the online news sources ecosystem, but we need more analysis to investigate which characteristics of the news outlets drive actual audience engagement [37]. Therefore, a more nuanced news source characterization is needed [1]. In this thesis, I utilize k-means clustering methods as an attempt of a more nuanced news source characterization, where I investigate whether any clusters of news sources naturally come up given online audience-based metrics as their features.

One method of analysis used in this thesis is clustering, specifically k-means, which aims to cluster entities by minimizing within-cluster sum of squared distances of the points in each cluster. Barros et al. has employed the same method on Reddit-based dataset to classify health content on Reddit in an unsupervised manner [4]. However, they utilize word embedding of posts instead of audience-metrics such as sharing frequency, upvote ratio, and number of comments for embedding as in this thesis. In this thesis, I investigate whether the same clustering algorithm but with embedding based on audience-metrics can yield insights regarding news source characterizations.

Other subreddit-based features have been used for clustering tasks as well. For example, Chandrasekharan et al. clusters subreddits based on their features of how likely those subreddits moderate comments posted in them, as well as based on their

specific norms [6]. Morrison and Hayes, on the other hand, use specific audience-based metrics as features to characterize the audience themselves based on "their popularity and role in initiating and sustaining communication in their communities" [28]. These features include the number of posts and comments the users engage in, the number of comments that received at least one reply, the number of subreddits the users engage in, and others. Similarly to Morrison and Hayes, this thesis utilizes features based on Reddit users" engagement on the platform. However, this thesis aggregates such features based on subreddits instead of treating them individually as Morrison and Hayes do. Differently to Morrison and Hayes, this thesis utilizes these features to cluster and gain insights about news sources instead of about the audiences themselves.

Chapter 4

Introducing the Datasets

In this section I will introduce the sources of data being used in this thesis. I use GDELT and Muck Rack to build my set of news sources and the Reddit Pushshift collection to extract features that demonstrate how these news sources are shared by their audience.

4.1 GDELT and Muck Rack to define "News Source"

As mentioned in Section 2.1, this thesis relies on the GDELT Project and the Muck Rack company to build a set of sites recognized as "news sources". Here, I will introduce both GDELT and Muck Rack as an organization and a company, respectively. I explain how I acquire data from each of them in Section 6.1.

4.1.1 GDELT

The GDELT Project aims to report all events happening all around the world by indexing and using NLP to process news. GDELT is a far-reaching multi-year project that has been used by multitudes of other researchers, and is freely accessible to everyone. GDELT was authored by Kaleev Leetaru and Philip Schrodt. Leetaru is a Media Fellow at the RealClearFoundation and a Senior Fellow at the George Washington University Center for Cyber and Homeland Security, where he also serves on

its Counterterrorism and Intelligence Task Force[1]. Leetaru still maintains the project and writes articles about its use cases as hosted on realclearpolitics.com[2]. Some of the most recent articles that Leetaru wrote using data from GDELT include an article about how the amount of coverage of Russia's invasion of Ukraine has been declining, less than two months since the start of the invasion[3]. Schrodt is a political scientist who is currently a senior research scientist at the consulting firm Parus Analytical Systems, after he left Pennsylvania State University where he was a professor. Schrodt created the Textual Analysis by Augmented Replacement Instructions (TABARI) software and co-developed Conflict and Meditation Event Observation (CAMEO) data coding framework, both of which were utilized in the making of GDELT [25].

Today, GDELT contains over 250M event records covering the entire world from 1979 until present, with new records being added every 15 minutes. As a project, GDELT aims to construct a single massive network that captures what is happening around the world, the context of the events, who is involved, and how the world is feeling about it, every single day[4]. GDELT is currently the largest open-access database on human society events[5].

GDELT records events happening all around the world in various structures including the "Global Knowledge Graph (GKG)" database, where the project compiles actors and places from every news report. For each of the news report recorded by GDELT, the GKG database also records its details such as the link of the news report; the domain name of the news source publishing the report; the numerical summaries included in the news report if available; themes; locations, organizations, and persons mentioned in the article; the tone of the article; image link; and video link.

Overall, GDELT has built its events record using more than 240,000 news sources originating from all around the world, including more than 13,500 English language sources. However, not all of these news sources are still active today, considering

[1]https://www.kalevleetaru.com/vita.pdf
[2]http://realclearpolitics.com
[3]https://www.realclearpolitics.com/video/2022/04/11/ukraine_is_already_fading_on_television_news.html
[4]https://www.gdeltproject.org/about.html#creation
[5]https://cloudplatform.googleblog.com/2014/05/worlds-largest-event-dataset-now-publicly-available-in-google-bigquery.html

that news sources undergo changes or even shut down over time, and that new news sources emerge.

As a source of data, GDELT has been used in many research projects, with most of them focusing on political-related events. For example, Qiao et al. utilize GDELT to predict social unrest events with hidden markov models [33]. On the other hand, Coniglio et al. utilized GDELT as part of their dataset to investigate "the effect of hosting refugees in camps on the occurrence of protests and social conflicts" where GDELT data is used to determine the frequency of protests and social conflicts [8]. Another use case is by Nejjari et al. where the authors employ graph theory to verify the assumptions that cyber threats are correlated to geopolitical events, by comparing graph isomorphism and structures of the cyber events graph and geopolitical events graph built using GDELT data [29].

It is worth mentioning that GDELT has also been critiqued in the past, especially regarding the concern of event duplications such as the kidnapping in Nigeria, as noted by Jenkins and Maher [20], and other researchers. However, this thesis is not concerned about the actual events and other labels reported by GDELT. Instead, this thesis looks into GDELT to only aggregate the set of news sources recognized by GDELT in their event reporting process. Therefore, even with a critiqued event labeling algorithm, GDELT is still a valid and important data source in this thesis's use case.

4.1.2 Muck Rack

As a company, Muck Rack claims to provide access to the "most accurate media database" where people can "monitor news as it breaks". Muck Rack's main consumer targets are other companies' Public Relations Managements (PRMs), as they can use Muck Rack to build relationships with various media outlets and analyze how their companies are represented by the media; and journalists, as they can showcase their portfolio and measure the impact of their published articles[1]. Unlike GDELT, Muck Rack is a paid service. However, it does have some resources open for the public to browse on their website, such as their media database, which is what this thesis uses.

As a data source, Muck Rack data is mainly used in journalism-related research.

GDELT and Muck Rack are two huge databases of news sources built for different purposes, but they both strive and claim to be complete media databases able to provide a thorough news coverage. Intersecting these two databases gives us an insight into the agreement between the private and public organizations regarding what counts as a "news source".

4.2 Reddit Pushshift Dataset

Pushshift is a project that collects and archives longitudinal social media data, including Reddit in addition to other platforms such as Tik Tok and Twitter. It is open to the public and is continuously updated. The Pushshift dataset has been used in various research projects of all scales, either through the raw data dumps, Google Big Query access, or its API. These projects include a study looking at discussions on skincare addition on Reddit [31], another study looking at posts discussing synthetic opioids in a retrospective observational study [5], a study focusing on language modeling and other NLP-related tasks [35, 22], COVID-19 related studies (including misinformation regarding the pandemic) [9, 30, 26], among others.

Even though Pushshift does provide easy access to its data through the Pushshift API[6], I chose to access their monthly data dumps directly[7] due to limited documentation, rate limitations, a few inaccuracies, and service outages every once in a while for the API. Even so, the Pushshift API does provide better service than the official Reddit API in various aspects, such as a five times higher rate limit and direct comparison to the texts of Reddit submissions and comments; so its API is still a viable alternative of getting datasets regarding Reddit activity.

Each monthly data dump is stored in Newline Delimited JSON format (NDJSON), with a decompressed size of about 130GB each file. To store and access such a big dataset, I keep the files compressed (9GB each) and utilize the zstandard Python

[6]https://github.com/pushshift/api
[7]https://files.pushshift.io/reddit/submissions/

```
{
    'author': 'elanglohablante9805',
    'author_created_utc': 1609519842,
    'author_patreon_flair': False,
    'created_utc': 1617235201,
    'edited': False,
    'id': 'mhj2hj',
    'num_comments': 2,
    'permalink': '/r/WriteStreakES/comments/mhj2hj/streak_90_ha_llegado_la_primavera/',
    'pinned': False,
    'pwls': None,
    'quarantine': False,
    'removed_by_category': None,
    'retrieved_utc': 1623447663,
    'selftext': 'Los pájaros están cantando, las hierbas verdes están brotando, y tengo alergi
as.  Esto es la temporada de las alergias.  Estornudo cada mañana cuando me despierto, y otra
vez si voy afuera.  Necesito tomar medicina cada día, pero no funciona tan bien.',
    'subreddit': 'WriteStreakES',
    'subreddit_id': 't5_2eamt5',
    'title': 'Streak 90: Ha llegado la primavera',
    'upvote_ratio': 1.0,
    'url': 'https://www.reddit.com/r/WriteStreakES/comments/mhj2hj/streak_90_ha_llegado_la_pri
mavera/',
    ...
}
```


Figure 4-1: A snapshot of some of the key-value pairs in one of the Python dictionaries representing a Reddit submission.

package, which allows me to stream the compressed data. At the time of writing this thesis, Pushshift data dates back to 2005 and the most recent dataset uploaded to Pushshift is for June 2021, which was uploaded in August 2021.

Each Reddit post in a monthly dump file is represented by a Python dictionary data structure with 82 keys, recording various features of a post, including most importantly the URLs contained in the submission, the text of the submission, subreddit name where it is submitted, subreddit subscribers, number of comments associated with the post, ratio of upvotes and downvotes, UTC timestamp of submission, and the UTC timestamp of when Pushshift adds the post to its database, among others. Figure 4-1 shows a snapshot of some of the key-value pairs in one of the Python dictionaries representing a Reddit submission.

On another note, there are two things that are worth mentioning: the fact that Pushshift treats some Reddit "users" as "subreddits", and that a lot of the links in the

submission files are `reddit.com`.

The first fact suggests that I should filter out these non-subreddits from my data of subreddits. Recall that subreddits are communities within Reddit that are created by an user and moderated by one or more users, while users are the actual reddit accounts that people sign up for. Specifically, subreddits are noted with an "r" notation on Reddit, while users are often noted with a "u" notation on Reddit. For example, the specific community on Reddit that discusses Pushshift is the `pushshift.io` subreddit, represented by the notation `/r/pushshift/`, which is accessible on reddit using the URL `https://www.reddit.com/r/pushshift/`. On the other hand, the user that creates Pushshift (who is also a moderator in `/r/pushshift`) is `/u/Stuck_in_the_Matrix`. However, due to the fact that users are allowed to post on their own home page as opposed to posting in a particular subreddit, Pushshift treats some users as subreddits, i.e. by associating a number of Reddit submissions as being posted in subreddits that are actually users' usernames. Since the main usage of Reddit is the utilization of its theme-specific subreddits for people to discuss and ask questions, very few users actually post in their own homepage because of the lack of traffic there by other users. Individual user's homepage is also not theme-specific and is only based on the preference of the user, therefore they do not add much information about the audience of these news sources. Thus, this thesis will ignore posts made on users' homepages.

Secondly, there are a lot of reddit.com URLs in the submission JSONs, which is due to the many inter-site sharing within Reddit, and because the "URL" field in the JSON files point to post itself if there are no outside URLs being shared. As explained in the About News and Who Produces It section, just like other news source aggregators and user generated content hosts, `reddit.com` will not be included in the analysis as a news source.

Chapter 5

Data Exploration

My first approach to tackle the task of characterizing news sources is by exploring what news sources are included in my set of news sources given my method of building this set of news sources, and exploring how they are shared on various different subreddits. I expect that various characteristics of news sources can be inferred from the patterns of how they are being shared. For example, I expect that we can infer the popularity of news sources based on the total number of their news links that are shared on Reddit, and in how many subreddits they are shared. I also expect that we can infer the audience broadness and appeal of some news sources by comparing their general popularity (i.e. the total number of news links shared from that domain) and their popularity across different subreddits.

5.1 Exploring News Sources

Intersecting news sources from GDELT and Muck Rack yields 42,477 news sources. Out of those news sources, there are 23,575 news sources that were shared on Reddit at least once in the span of January - June 2021. Out of these, only 8,781 news sources were consistently shared at least once per month in the same time span. In the "Methods and Results" section, I will go into details about the ways I filter these sites to only include news sources that are not sites of news aggregators and hosts of user-generated content. In the same section, I also explain how I further subset the

set of news sources to filter out news sources that are shared on Reddit extremely infrequently thus lack representation. In this section, I will go over the statistics of news sharing on Reddit as a whole.

As shown in Table 5.1, there are more than 30 million monthly submissions on Reddit, and about 60% of those submissions contain at least one URL. Among the submissions that contain at least one URL each, 25% contain links to news sources and user generated content such as blogs and podcasts. Our set of news sources makes up about 7% of the links being shared on Reddit posts.

Month	Total submissions	Submissions with link(s) excluding 'reddit.com'	Submissions with news source & user generated content links	Submissions with news sources as defined here
Jan	32,704,571	19,550,221	4,402,183	1,358,227
Feb	31,147,947	18,296,269	4,186,966	1,248,924
Mar	33,006,103	19,494,992	4,639,876	1,387,165
Apr	31,616,206	18,505,024	4,639,876	1,387,165
May	36,310,673	21,850,116	4,670,423	1,314,159
Jun	34,118,481	20,050,925	4,488,769	1,022,494

Table 5.1: Among the submissions that contain at least one URL each, 25% contain links to news sources and user generated content such as blogs and podcasts. Our set of news sources makes up about 7% of the links being shared on Reddit posts.

Each of the sites in the set of news sources are not all equally popular, as some are shared more than the others, regardless of the number of subreddits they are shared in. 'imgur.com' for example, was shared more than 9 million times, and 'youtube.com' more than 4 million times. Many other news aggregators and user-generated content hosts are similarly shared with high frequency. It is interesting to note that these sites are extremely popular on Reddit, however, as discussed previously, I will disregard these sites.

The list of sites that are shared the most on Reddit is summarized in Table 5.2.

As part of the data cleaning process, as will be explained later, I manually went through this ranked list of news sources and filtered out heavily-shared sites that are not news sources, but are news aggregators or user generated content hosts instead, which have no editorial oversight. Although these platforms are interesting to learn about, this thesis focuses on actual news sources. I filtered out 133 such sites. The complete list of these 133 sites is in Table B.1 in Appendix B.

Excluding news aggregators and user-generated content hosts, we are left with the

Sites	# times they are shared (regardless of the subreddits)	Proportion based on total # times all sources are shared (%)
imgur.com	9957691	35.3%
youtube.com	4219365	14.9%
twitter.com	1888497	6.7%
t.me	763911	2.7%
google.com	678895	2.4%
mlb.com	318979	1.1%
blogspot.com	317397	1.1%
spotify.com	317203	1.1%
theguardian.com	226634	0.8%
steamcommunity.com	185928	0.6%
instagram.com	168276	0.6%
wikipedia.org	164672	0.6%
cnn.com	161038	0.6%
amazon.com	149672	0.5%
facebook.com	146110	0.5%
github.com	145580	0.5%
nytimes.com	144988	0.5%
foxnews.com	141893	0.5%
medium.com	130885	0.5%
soundcloud.com	126885	0.4%

Table 5.2: Top 20 sites (not necessarily news sources) that are shared the most on Reddit.

actual news sources. Amongst them, ‘mlb.com‘, which is the official site associated with Major League Baseball, is the most shared news source as it was shared 318,948 times in total in the timeframe of interest, encompassing 4.5% of the total news sharing frequency of over 7M. The list of the top 20 news sources that are shared the most is summarized in Table 5.3.

In terms of statistics, the mean sharing frequency of news sources is 300 with median 10, representing an extremely right-skewed distribution.

One natural question to ask is if news sources that are in total shared the most, are the same news sources that are shared in the most subreddits (i.e. popular across

News sources	# times they are shared (regardless of the subreddits)	Proportion based on total # times all sources are shared (%)
mlb.com	318948	4.5%
theguardian.com	225921	3.2%
cnn.com	160451	2.3%
nytimes.com	144252	2.0%
foxnews.com	141796	2.0%
thehindu.com	88410	1.2%
bbc.co.uk	87890	1.2%
thestar.com	87662	1.2%
reuters.com	85021	1.2%
nypost.com	81594	1.1%
cnbc.com	65440	0.9%
thehill.com	63500	0.9%
indiatimes.com	63034	0.8%
washingtontimes.com	59941	0.8%
breitbart.com	57660	0.8%
usatoday.com	54144	0.7%
scmp.com	52531	0.7%
cbc.ca	52372	0.7%
apnews.com	48543	0.7%
nbcnews.com	48378	0.7%

Table 5.3: Top 20 news sources that are shared the most on Reddit.

different subreddits). Now that we have looked at the total sharing frequency of each of the news sources, let's look at the number of subreddits in which these news sources are shared.

News aggregator sites and user-generated content hosts are popular across different subreddits, as expected, since these sites are rarely theme-specific. In the analysis below, I will discard them and only consider news sources.

As shown in left hand side graph of Figure 5-1, a lot of news sources are each shared in 100 or less different subreddits, but there are some that are shared in up to 1000 different subreddits. In fact, about 4.5K (almost 20%) out of the total 23K news sources are only shared in one subreddit each. From the graph in the right

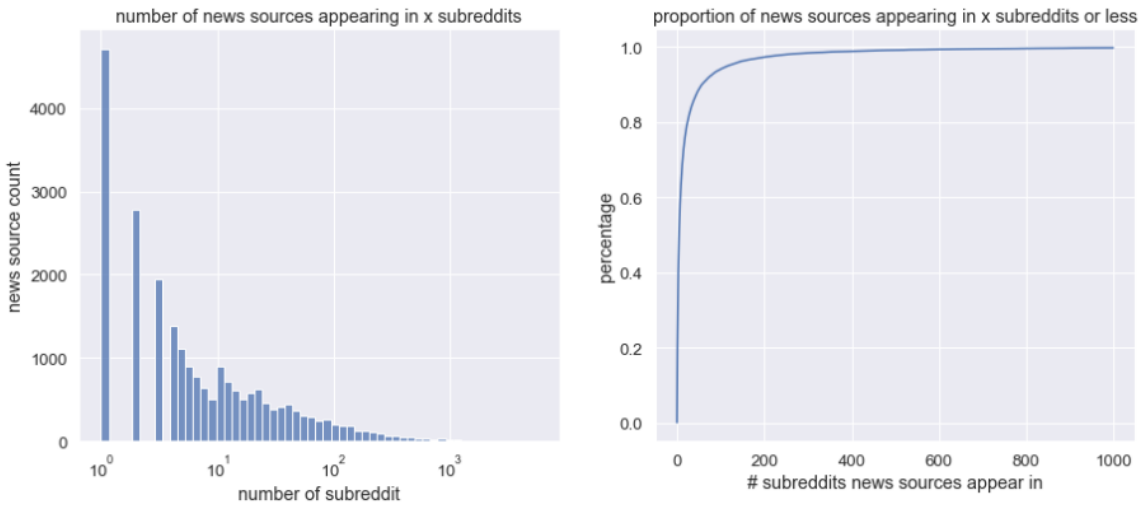

Figure 5-1: Figure 5a shows that a lot of news sources are each shared in 100 or less different subreddits, but there are some that are shared in upto 1000 different subreddits. Figure 5b shows the same information, that most (98.8%) news sources appear in 400 subreddits or less.

hand side graph of Figure 5-1, we can also see that almost all news sources are only shared in a maximum of about 400 subreddits. In fact, the mean of the number of subreddits a news source is shared in is 31, while the median is 5, representing the extremely right-skewed distribution. The fact that about 50% of the set of over 23K news sources were only mentioned in less than 5 subreddits suggests that it might be beneficial to also set a lower bound threshold of minimum number of subreddits as I try to cluster these news sources later using the subreddits as the features.

Next, let's look into what news sources are most popular across different subreddits, ranked by how many subreddits contain posts that share these news sources' links. The top 20 news sources are shown in Table 5.4.

The following news sources are in the list of top 20 news sources with the highest number of links shared, but not in the list of top 20 news sources with the most subreddits they are shared in: mlb.com, foxnews.com, thehindu.com, thestar.com, thehill.com, breitbart.com, and cbc.ca. In fact, thestar.com, thehindu.com, breitbart.com and mlb.com only place in 108, 209, 215, and 522 rank respectively in terms of the number of unique subreddits they are shared in. This suggests that these 4 news sources are very heavily shared in a small subset of subreddits.

News sources	# subreddits they are shared in
theguardian.com	5617
nytimes.com	5438
cnn.com	5005
bbc.com	4241
reuters.com	3906
washingtonpost.com	3692
forbes.com	3488
cnbc.com	3357
bloomberg.com	3269
apnews.com	3084
businessinsider.com	3026
bbc.co.uk	2971
npr.org	2906
nbcnews.com	2815
independent.co.uk	2728
msn.com	2699
dailymail.co.uk	2693
vice.com	2581
nypost.com	2441
wsj.com	2409

Table 5.4: Top 20 news sources ranked by the number of subreddits they are shared in.

Comparing these two measures in this way can give us insights into what news sources are heavily shared by particular subreddits as well as what those subreddits are. This insight might be easily explained, for example in the case of mlb.com, it is very niche because it covers a very specific topic that is baseball, so it is only appealing to, and thus shared by, a small subset of the Reddit community. However, this insight is more interesting for breitbart.com, for example. The news website breitbart.com is a far-right news outlet that notoriously spreads misinformation, most recently about Covid-19, and in general fails to report news factually. The heavy sharing frequency of this news outlet in a small set of subreddit might play a factor in creating and sustaining an echo chamber of misinformation spread. This could be one way of characterizing news sources, in particular to find out which news sources are most heavily used to spread certain ideology.

Visually, Figure 5-2 shows the distribution of news sources based on their total mention count as well as the number of subreddits in which they are mentioned. The figure itself is divided into four sections labeled as quadrant 1, 2, 3, and 4, by drawing the green horizontal line Mention count = 1000, and the orange horizontal line Subreddit count = 1000. Note that these are high thresholds considering that, as mentioned before, overall news source mention count has median of 10 and mean 272, while the number of subreddits in which the sources are shared has median of 5 and mean 31.

The first quadrant contains news sources that have the highest mention count of more than 1000 mentions, and are shared in proportionally many different subreddits (>1000 subreddits). These news sources are popular as they are shared many times, and also have a broad appeal as they are shared in a lot of different subreddits. In total there are 68 such news sources (0.29% of the total set of news sources). Some examples are buzzfeed.com, thehill.com, nytimes.com, dw.com, cbc.ca, and others.

The second quadrant contains news sources that have the highest mention count of more than 1000 mentions, but are mentioned in comparably fewer subreddits. These are news sources that are popular but are more niche. In other words, they are heavily shared by only particular communities. In total, there are 734 such news sources (3.13% of the total set of news sources). Some examples are mlb.com (shared 318,947 times in only 237 subreddits), breitbart.com (shared 57,659 times in 496 subreddits), nhl.com (shared 38,763 times in 148 subreddits), oann.com (shared 27,106 times in 144 subreddits), hotnews.ro (shared 13,195 times in only 22 subreddits), and others. Theme or non-English language specific news sources such as mlb.com, nhl.com, and hotnews.ro are intuitively niche because only particular sets of communities understand and are invested in those types of news sources. However, knowing that some far-right leaning news sources such as breitbart.com and oann.com (also known as One America News Network) gives us an insight into how the perspectives of these news sources are shared online. This helps us characterize these types of news sources as niche news sources.

The third quadrant contains news sources whose sharing frequency and count

of subreddits are not of special interest. These news sources are neither notably popular nor have niche audiences. There are 22,615 news sources in this quadrant (encompassing 96.6% of the total news sources). Note that there are only news sources in the top left part of quadrant 3. This makes sense since the number of subreddits in which a news source is mentioned cannot be higher than the total mention of that news source (e.g. if a news source is only mentioned in 10 subreddits, then it must have 10 or less total mention counts). The fourth quadrant is empty for the same reason.

In conclusion, by looking at the measures of mention count and subreddit count for the news sources, we are able to gain insights about which news sources are more broadly popular and which are more niche. This type of characterization is very valuable for understanding how news sources are consumed by online audiences, yet the characterization process was able to be done simply by comparing the two metrics above.

5.2 Exploring Subreddits

Now that we have a good understanding of how news sources are shared in various subreddits, let's look into the subreddits themselves.

Including both news aggregators and user-generated content hosts in addition to news sources, there are 131,633 different subreddits in which people share at least one link within the time frame of January to June 2021. As a side note, this number (131,633) makes sense because according to Reddit[1], it only has 100k+ active communities out of the millions of subreddits that Reddit has in total. Out of these, only 45,777 subreddits were consistently used at least once per month in our time frame of interest.

Excluding posts that only share links to news aggregators and user-generated content hosts, there are 43,800 subreddits in total, in which people share at least one news source link. Only a quarter of these were consistently used to share news links at least once per month within January - June 2021 time frame.

To understand the popularity and rate of usage of the subreddits in my dataset, I look into the number of their subscribers (equivalent to members) as well as the activity with those subreddits in terms of the number of posts. For the analysis below, I exclude news source aggregator sites as well as hosts of user-generated content.

5.2.1 Subreddits Subscribers

First, we look into the number of subscribers of each of these subreddits. Out of the 43,800 subreddits, there are 20,637 subreddits that have subscribers' information.

The distribution of the number of subscribers is extremely right-skewed, so I use log (base 10) to visualize the logged-subscribers distribution, as follows.

As we see from Figure 5-3, about half of the subreddits have about 0-1,000 subscribers. The other half have about 10,000 to 10,000,000 subscribers.

Table 5.5 presents top twenty subreddits based on the number of their subscribers.

1. funny	6. worldnews	11. news	16. food
2. gaming	7. Music	12. Showerthoughts	17. Jokes
3. aww	8. videos	13. IAmA	18. explainlikeimfive
4. pics	9. movies	14. EarthPorn	19. books
5. science	10. todayIlearned	15. askscience	20. LifeProTips

Table 5.5: Top 20 subreddits ranked based on the number of their subscribers.

Note that these subscribers are not unique, as one person can subscribe to several subreddits (but one can only subscribe to each subreddit once). In future work, more data should be analyzed to rerun this analysis while taking into account the difference between users.

From the table above we see that most subreddits that have a lot of subscribers are very general and are not specifically associated with any real-life community. The subreddit r/pics which is a subreddit for pictures, for example, do not specifically appeal to any particular real-life communities and thus do not contain any more audience-specific information than the subreddits with less subscribers. Thus, when considering the subreddits to use as features of news sources in the process of characterizing them, filtering by the number of their subscribers might not be such a good

idea. This is a useful insight for when I consider ways to reduce the dimensions of news sources' vector representations using the subreddits in which they are shared in, as the features.

5.2.2 Subreddits Activity

The second measure of the popularity and traffic of subreddits, is the activity within them. Here, I define 'activity' as the number of posts made within those subreddits.

Based on the number of submissions made to all subreddits during January to June 2021, Table 5.6 presents the top 20 subreddits with the most activities.

1. AutoNewspaper	6. COVID_CANADA	11. NoFilterNews	16. CertifiedNews
2. politics	7. worldnews	12. FakeCollegeFootball	17. WrestlingBreakingNews
3. TheNewsFeed	8. Conservative	13. nofeenews	18. FOXauto
4. news	9. THEHINDUauto	14. TORONTOSTARauto	19. NewsfeedForWork
5. TrendingQuickTVNews	10. niuz	15. trendandstyle	20. Coronavirus

Table 5.6: Top 20 subreddits with the most submissions posted in them.

This list of top 20 subreddits based on the number of submissions in them, is rather consistent in the time frame of January to June 2021. Since we are focusing on news source links, most of the submissions are made in news-related subreddits, which is expected. Note that there are a couple of subreddits that are aimed to host posts posted by bots. r/AutoNewsPaper for example, claims to "provide diverse uncensored news via individual subreddits and a combined subreddit of the most circulated English language news sources in the world," with posts made by automated user (bot) which acquires the news articles from news source RSS directly. In the final analysis, these subreddits are discarded.

In addition to submission counts, I also investigate the set of subreddits with the highest number of different news sources mentioned in them. Table 5.7 records the list of top 20 subreddits with the most news source entities mentioned in them.

As seen in Table 5.7, the subreddits with the highest number of mentioned news sources are generally news-themed, which makes sense. It is interesting to note that there is very little overlap between the top 20 subreddits based on how many unique news source entities are shared in them, and the top 20 subreddits based on the

1. COVID_CANADA	6. worldnews	11. wallstreetbets	16. CryptoCurrency
2. news	7. nottheonion	12. tomorrowsworld	17. technology
3. politics	8. prisons	13. autotldr	18. europe
4. todayilearned	9. NoFilterNews	14. Conservative	19. electionReformNews
5. Coronavirus	10. conspiracy	15. NoNewNormal	20. HumanTraffickingNews

Table 5.7: Top 20 subreddits with the highest number of unique news source entities shared in them.

number of submissions made in them in total.

In general, a higher number of news links shared in a subreddit means there are a higher number of unique news sources mentioned in that subreddit. In other words, there is a positive correlation (r=0.55) between the number of unique news sources mentioned in a subreddit and the total number of news links shared in that subreddit. This positive correlation is shown in Figure 5-4. However, there are some subreddits where this is not the case, for example subreddits that are moderated by bots where only articles from specific news sources are shared. Some examples are r/BBCauto, r/CHICAGOSUNauto, r/CNET_ALL_RSS, and others. These subreddits are removed in my analysis.

Among the 131,633 different subreddits in which people share at least one news link within the time frame of interest, 62,683 subreddits (47.6% of total subreddits) only have 1 news link mentioned in it. This suggests that I should not use this big set of subreddits in its entirety for the clustering task, instead I should filter these subreddits to only include a set of those with the most complete information regarding our set of news sources. One way to approach this is by employing PCA (Principal Component Analysis) to form a set of linear combinations of these subreddits and choosing a subset that explains the most variance within these subreddits. This is implemented in the "Pre-processing data" section.

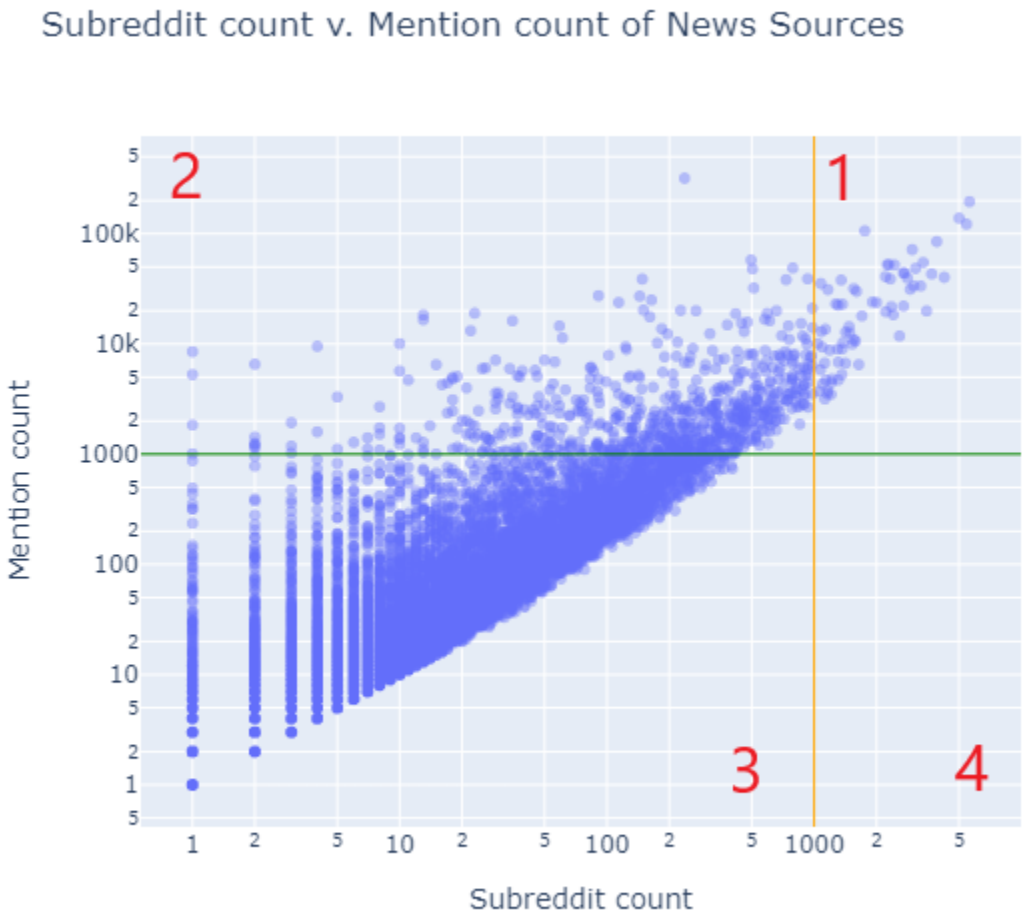

Figure 5-2: Subreddit count v. Mention count of News sources. Quadrant numbers are labeled in red. News sources in the first quadrant are both mentioned very frequently and in many subreddits (broad appeal), for example theguardian.com, cnn.com, and nytimes.com. News sources in the second quadrant are mentioned very frequently but in fewer subreddits (more niche), for example mlb.com, thehindu.com, and breitbart.com.

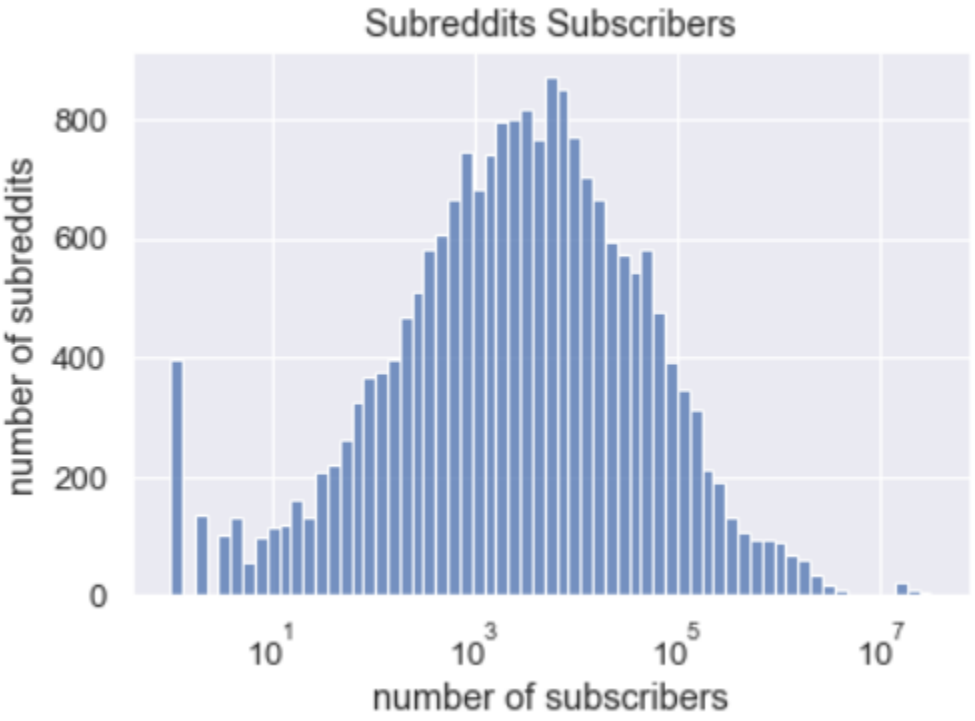

Figure 5-3: Subreddits subscribers. More than half of the subreddits in our dataset have about 1000 or less subscribers.

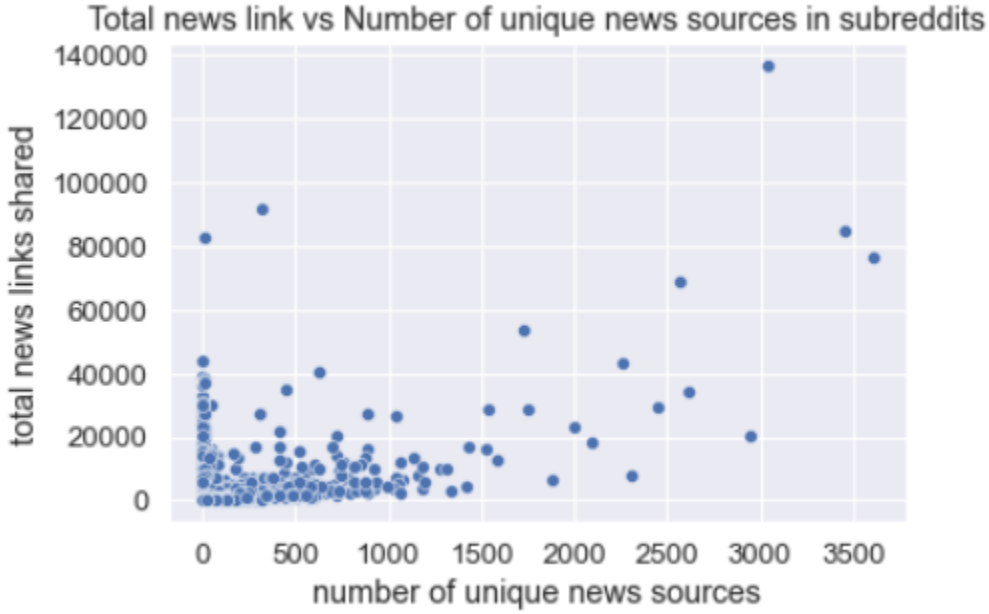

Figure 5-4: There is a positive correlation between the number of unique news sources and total news links shared in each subreddit (r=0.55). Some outliers, shown in the bottom left corner, are subreddits that are moderated by bots to only share news from specific news sources, such as r/BBCauto among others.

Chapter 6

Methods for Data Representation

In this section I will broadly go over the steps I take to process my dataset and utilize it to infer news source characterizations, then talk more deeply about it in the subsections that follow.

Recall that my goal is to investigate news source characterizations based on their online audience. To do this, I will first represent each news source using the statistics of the consumption by their online audience. Then, using these representations as features, I will first investigate if clustering the news sources yields results in a representation that places 'similar' news sources together. I define 'similar' as having one or more identical news characteristics, such as media type, language, origin, themes, reliability, etc. Using some news sources that are labeled based on these characteristics, I will then make note of the characteristic similarities that are most prominent among news sources that are represented closer together. If I find that it is the case that similar news sources are indeed placed closer together using this method of representation, then I reason that we can characterize unlabeled news sources using this kind of representation, by inferring characteristics of the news sources that are close neighbors of those unlabelled news sources. In this thesis, the label of some characteristics for some news sources is obtained using Wikidata[1] which is a free knowledge based of 97M data items, and Media Bias Fact Check[2], which is a fact-

[1]https://www.wikidata.org/wiki/Wikidata:Main_Page
[2]https://mediabiasfactcheck.com/

checking website that rates the reliability and factuality of news sources that is open to the public.

The second way that I investigate the usefulness of audience-based metrics in news source characterization is by building classification and regression models, using the audience-based metrics to predict some news source characteristics. If good-performing models are found, then I reason that we can characterize unlabelled news sources by using their audience-based representations and these classification and regression models.

I represent my dataset of news sources using the number of times they are shared in users' posts in various subreddits (i.e. their sharing frequency), as well as that sharing frequency weighted by the number of comments and upvotes ratio that those posts have. In other words, I represent news source entities using the statistics of their presence in various subreddits as their features.

Before building vectors representing these news sources, I first filter out popular sites that are not news sources, but instead are news aggregators or hosts of users' generated contents. Recall that there are a lot more subreddits than there are news sources. Therefore, due to the extremely high feature dimension of the news source entities, I also filter out subreddits that only contain very few news sources shared in them.

However, heavily filtering out subreddits based only on the sharing volume in them is not a good idea because we will only be left with very general subreddits that will not bring very much information about the characteristics of the news sources. Therefore, after scaling the data, I also employ the Principle Component Analysis (PCA) method for dimensionality reduction, with the hope of reducing the number of features while still retaining most of the information they bring.

6.1 Building news source dataset

Recall that I built my dataset of news sources by intersecting the set of news sources according to GDELT and the set of news sources according to Muck Rack. Here I

will describe how I went about getting these two sets of news sources.

To acquire the set of news sources recognized by GDELT, I focused on its Global Knowledge Graph (GKG) table as hosted on Google Big Query[3]. Although there is a lot of information included in the table, I only extracted the information about the domain name of the news source. To get as many active news sources as possible from GDELT, I collected the domain name of the news sources that GDELT has used in the past 5 years, 2016 - 2021. In total there are 205,994 unique news source domain names that I got from GDELT.

Muck Rack, on the other hand, does not host their data in such an easy-to-query database. To acquire the set of news sources recognized by Muck Rack, I utilize their publicly available table of news sources hosted on their website[4] using Python's selenium library. I extracted data from Muck Rack in two iterations. The first time of scraping was in September 2021, where I acquired 67,782 news sources from Muck Rack. Since Muck Rack keeps on updating their database thus changing the total news outlets that they have, I went to the site again in January 2022, where I acquired 450,590 total news sources. This is a dramatic increase in the number of news sources, but it is not the case that the number of global news sources more than quadrupled in the timespan of four months. In fact, I found that most of the new sites that were added were links to individual podcasts, including those hosted on anchor.fm and spreaker.com. However, just like other user-generated contents such as blog posts, these podcasts are filtered out in a later stage of data processing, since, by the definition used by this thesis, these user-generated contents and the sites that host them do not count as news sources.

I built the final set of news outlets recognized by Muck Rack by combining the two sets of news sources extracted in September 2021 and January 2022 via the union operation, resulting in 457,882 total news sources.

To build my final set of news sources, I intersected the set of news sources from GDELT and the set of news sources from Muck Rack. Their intersection results in

[3]https://cloudplatform.googleblog.com/2014/05/worlds-largest-event-dataset-now-publicly-available-in-google-bigquery.html

[4]https://muckrack.com/media-outlets

42,477 news sources, which comprises about 20% of the total news sources appearing in GDELT, and 9% of the total news sources appearing in Muck Rack. As shown in Table 6.1, it is interesting to note that these two organizations have very little agreement on what counts as news sources. GDELT claims to cover most if not all events happening all around the world, thus utilizing most if not all global news sources. Similarly, Muck Rack claims to provide a complete media outlets database. Yet, the two agree very little on what counts as news sources. This emphasizes the need for us to gain further understanding of news sources. This thesis contributes to this purpose by trying to understand the possibilities of deriving insights about news sources based on their online audiences.

Data source	Number of news sources
GDELT	205,994
Muck Rack	457,882
GDELT ∩ Muck Rack	*42,477*

Table 6.1: Summary of the number of news sources recognized by GDELT and Muck Rack, and the little overlap between the two.

After intersecting the set of news sources acquired from GDELT and Muck Rack, I stored this set of news sources and later went on Pushshift Reddit database to see how each of them are shared in various subreddits. For ease of reference, I will refer to this set of 42,477 news sources as 'gm_intersection'.

6.2 Building News Sources Representation

Since my goal is to investigate news source characterizations based on their audience, I will represent each news source using the statistics of how they are shared by their online audience on Reddit. Just as the usual dataset setting in data analysis or Machine Learning framework is represented using various features, my dataset of news sources is also represented using their sharing statistics on Reddit as their features. In this section I will explain how I acquire the sharing statistics of news sources from Reddit and how I process and use them as features of my set of news sources.

6.2.1 Streaming and processing Reddit zst data

I acquired the dataset of Reddit submissions in January - June 2021 from the Pushshift database by manually downloading the zst-formatted files, totalling in around 50GB compressed size. Decompressing these files would require about 600 GB storage. Due to the prohibitively large uncompressed size of these files, I kept them compressed and utilized the zstandard Python package to stream the submission data in each of the files.

Figure 6-1: An example of Reddit submission. The highlighted parts are what I extract from the NDJSON files representing this submission. Note that the current version of Reddit only shows the total number of votes received (weighted by +1 for upvotes and -1 for downvotes). However, the Pushshift dataset contains the actual upvote ratio (number of upvotes divided by total votes), as does the old version of Reddit (old.reddit.com).

Recall from Section 4.2 that each of these files contains Python dictionaries representing submissions. Figure 6-1 shows an example of Reddit submission with the parts that I extract highlighted in yellow. Each of these submission dictionaries are associated with a subreddit and have various details recorded in the dictionary. I then utilized regular expressions to extract URLs from each of the submissions made on Reddit during the six months. Next, for each of the extracted URLs, I extracted the subdomain, domain, and suffix to form the host or domain name of the URLs. For each domain, I then checked if it is contained in my set of news sources gm_intersection. If it is, I then updated my records about the sharing frequency of this particular domain in the particular subreddit that the submission dictionary

is associated with. Within the same dictionary, there is also information about the upvote ratio and the number of comments that the post receives. Thus, in addition to keeping track of the sharing frequency of this particular domain in this particular subreddit, I also keep track of the same frequency weighted by the upvote ratio, as well as the same frequency weighted by the number of comments.

6.2.2 Data Representation

There are two distinct yet very connected entities in my dataset: news sources and subreddits. The connection between these two entities are very clear, which is that various news sources are mentioned in various subreddits thus forming connections between them. My goal is to then learn the connections within the set of news sources. In other words, I will try to infer the relationship between news sources based on how they are connected to the other set which is the subreddits.

An intuitive method of going about this is by representing the relationship between the two sets in a matrix form, let's call this matrix m, where the rows will represent each of the news sources, and the columns representing each of the subreddits. The actual values in the matrix at position $m[i][j]$ will then represent the weight of the relationship between the news source at row i and the subreddit at row j. The most straightforward way of defining the values in the matrix would be by using binarization, where we only use the value $m[i][j] \in 0, 1$, where $m[i][j] = 1$ if news source i was mentioned at all in subreddit j, and $m[i][j] = 0$ otherwise. However, representing news sources based on their Reddit audience in this way discards a lot of useful information such as the different frequency of news source mentions in the submissions in each of these subreddits, as well as the engagements with those submissions such as upvote ratio and number of comments. So, I employ a more expansive matrix representation that keeps track of three news source engagement statistics for each subreddit: sharing frequency, upvote-based weighted frequency, and comment-based weighted frequency. Note that comment-based weighted frequency is equivalent to the total number of comments that posts sharing news source links receive.

In order to keep track of these three statistics, instead of using only one column to represent each subreddit, I use three columns instead. The first column of a subreddit will record the raw number of sharing frequency of each news source, the second column will record the sum of the upvote ratio that each submission sharing each news source gets, while third column will record the sum of the number of comments that each submission sharing each news source gets. Each row, on the other hand, still represents one news source each. An example of such representation is shown in Table 6.2, for nytimes.com, with an example representation based on r/investing which is one of the subreddits being considered.

News Source	...	Sharing freq. in r/investing	Total upvote ratio in r/investing	Total comments in r/investing	...
...	...	...	...	...	...
nytimes.com	...	1392	121.79	7296	...
...	...	...	...	...	...

Table 6.2: An example of a news source representation. The example news source is nytimes.com, with an example representation based on r/investing which is one of the subreddits being considered. In a later section, this matrix will be further processed by scaling and dimensionality reduction.

6.2.3 Filtering Subreddits

There are three reasons why filtering subreddits is a good idea in this case. First, there are a lot more subreddits than news sources in the dataset, in other words there are a lot more feature columns than the data points (rows). This leads to what is commonly known as the "curse of dimensionality" in Machine Learning, where there are significantly more features than data points, which leads to poor-performing models[5]. Since I also represent each subreddit in three columns, the gap between the number of news sources and subreddits becomes even wider. Filtering out some subreddits will help in reducing the dimensionality of the dataset.

Second, as discussed previously in the Section 5.2, there are a lot of subreddits that do not have a lot of news source links shared in them thus contributing little to

[5]https://en.wikipedia.org/wiki/Curse_of_dimensionality

no information. Additionally, there are also a lot of subreddits that have very few news sources mentioned in them. As mentioned in Section 5.2.2, there are a lot of subreddits with bots as their moderators and posters, where these bots only share news links from particular news sources in an automated fashion. Since these kinds of subreddits do not bring in any information about news source audiences, I also exclude them from my analysis. In particular, I only included subreddits that have at least 10 unique news sources mentioned in them and at least 100 news links shared in them in the time span of six months (January - June 2021). Filtering subreddits in this way results in the final set of 2,189 subreddits. Even though this is a rather small subset of the original set of subreddits, the low threshold used to filter these subreddits assures us that we are not losing too much information.

Third, memory and computational power. My computer had trouble storing and processing the dataset with no or minimal subreddit filtering. Due to computing power and the limited time I have to work on my thesis, working with a smaller (yet not much less informative) subset of my dataset is a good choice. Scaling the method to cover the full set of the dataset, and even covering longer periods of time, is an interesting subject of future research.

All in all, the final set of 2,189 subreddits make up 2189 x 3 = 6,567 features.

6.2.4 Filtering News Sources

As discussed in Section 2.1, some of the sites that are included in our original set of news domains are not news sources, but instead are news aggregators and hosts of user-generated content instead. These two groups of sites are filtered out of our set of news sources.

I also have to consider the news sources that are not shared frequently enough in our dataset, which would be hard to characterize. For example, 35% of the news sources in our set of news sources were only shared less than 5 times. There are two possibilities of the low rate of sharing of these news sources. First possibility, these domains might not be valid domains and are included in the dataset due to scraping errors. Second, these domains are valid domains and are just shared very infrequently.

To confirm that this big proportion of domains are indeed valid, as opposed to invalid URLs resulting from problematic URL extraction, I randomly sampled 10% of these sites and investigated the HTTP response they return. Using the requests library in Python, I found that about 77% of these sites are valid (with 200 response codes). Another 5% returns SSL Error which indicates sites that are perceived to be unsafe (but they do exist). In total, 82% of this sample are valid sites, though some unsafe. The other 10% return 403 response codes, 0.4% return 404 response codes, and the remaining 7.6% return various other responses. Thus, scraping error is unlikely to be a problem, and it is simply the case that a big subset of my set of news sources were very infrequently shared on Reddit.

It is also hard to characterize these news sources that are shared in very few subreddits because there is not enough data about them. Therefore, I set a threshold that a news source must be shared at least 50 times within the first six months of 2021 in order to be included for analysis. All news sources that were shared less than this are excluded from my analysis. This filtering process brings down the number of news sources in my analysis to 2,647 news sources.

In total, 2,647 news sources represented by 2,189 subreddits form a matrix with 2,647 rows and 6,567 feature columns.

6.2.5 Scaling and PCA Dimensionality Reduction

Note that even after the filtering process, we still have a very large number of subreddits compared to the number of news sources, thus potentially preserving the "curse of dimensionality". The next step in data processing is to employ Principle Component Analysis for dimensionality reduction. However, before employing this dimensionality reduction technique, I first scale my data.

Scaling is an important part of data processing in common data analysis or machine learning pipeline, if we have the right reasoning in doing it. In this case, I found that not standardizing my data results in poorer clustering and classification performance in inferring news source characteristics. Using clustering, for example, unscaled data yields in clustering popular news sources together, and not much else.

However, I hope to be able to extract other characteristics of news sources and represent them in their embedding space such that similar news sources are closer together, so we can infer characteristics (thus characterize) news sources based on their close neighbors. I find that scaling yields a better result.

For further dimensionality reduction, differently from the previous step, in this step I do not manually pick which subreddits to include as features. Instead, PCA builds a new feature set using linear combinations of the original feature set. The members of new feature sets are called the principal components.

There are two versions of PCA that I consider: a regular PCA, which is a LAPACK implementation of the full Singular Value Decomposition (SVD); and a randomized truncated SVD by the method of Halko et al. 2009 [6]. The two methods are similar, with the exception that the second method does not center the data as part of its preprocessing, among other differences. I consider the second method because it works well with sparse data, which my dataset is. Recall that the first objective is to cluster the news sources using their lower dimensional vector representations. Using various metrics to evaluate the clustering results such as silhouette coefficient, Calinski-Harabasz Index, and Davies-Bouldin Index[7], I found that the first version of PCA, which is the regular PCA, results in a slightly better clustering results, given that I first standardize the data, using the StandardScale() function from sklearn library. I then calculate the percentage variance explained by the principal components and summarize the result in Figure 6-2.

After applying PCA, I found that the first 300 principal components explain most of the variance of the subreddit-based features. To be exact, they explain 0.86 of the features variance. Thus, the final representation of the gm_intersection news sources is made of the 300 principal components as their features.

[6]https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html
[7]https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation

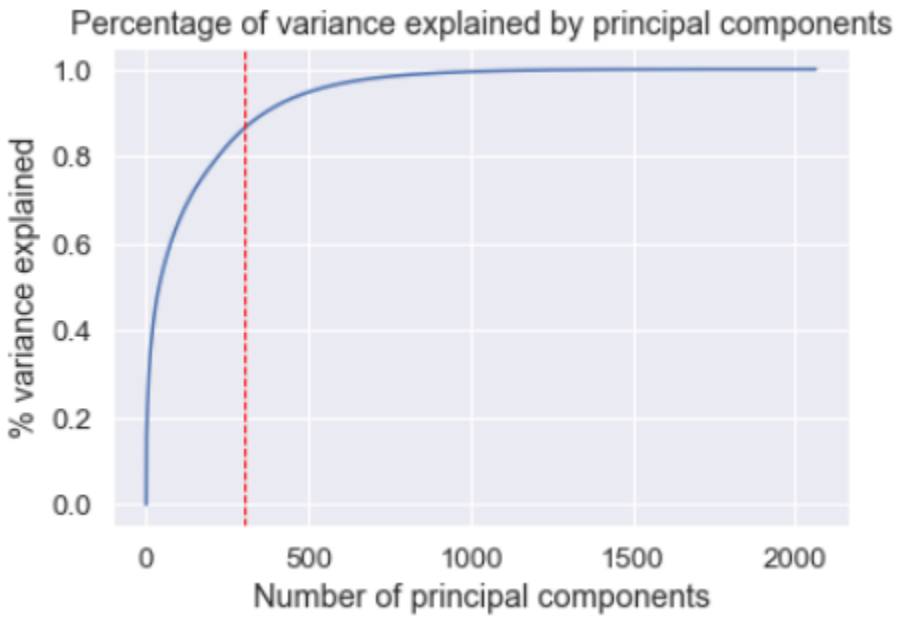

Figure 6-2: The figure above shows the cumulative percentage of variance explained by a set number of principal components derived from the subreddit-based features. The red line shows that the first 300 principal components explain 0.86 of the features variance).

Chapter 7

Methods for Data Analysis

As shown in Section 5.1, a simple comparison between mention count and subreddit count of news source sharing leads to an interesting news source characterization in terms of their popularity and broadness of appeal. Now that I have represented the gm_intersection news sources using their sharing metrics on various subreddits, in this section I will analyze whether this type of news source representation can be utilized to draw insights about news source characterizations.

To do this analysis, I first visualize the 300-dimensional representation of news sources down to 3-dimensional then visually investigate whether similar news sources are close together in this representational space. I then attempt to employ both clustering and classification methods to investigate whether any natural clusters of news sources emerge in this representational space, or whether we would be able to infer some specific characteristics of news sources (such as reliability, country of origin, political leaning, and media type) using the 300-dimensional representation.

In the following sections I will go over the three methods.

7.1 Visual Investigation

Visualizing news sources using all of their 300 features for representation is difficult if not impossible due to the high feature dimensionality. For clarity of visualizations, I project down the 300-dimensional representation to 3-dimensional representations

using T-distributed Stochastic Neighbor Embedding (t-SNE) as implemented in the Python's sklearn library[1]. One might wonder why I don't simply use the first three principal components that result from PCA for 3-dimensional representations of the news sources. The reason for this is that the first three principal components only explain less than 21% of the variance of all features, thus not ideal. On the other hand, simply using t-SNE, without PCA, to project down more than 6,000 features to 3 features is very expensive computationally and is generally not recommended. It is recommended to first use PCA to reduce the very high dimension to a reasonable amount then use t-SNE to speed up computation and suppress noise, just as it is done in this thesis.

As an overview, t-SNE is a stochastic algorithm that is commonly used to visualize high-dimensional data using a non-linear dimensional reduction method. It is based on the Stochastic Neighbor Embedding (SNE) algorithm developed by Hinton and Roweis in 2002, with the t-distributed variant proposed by van der Maaten in 2008. In summary, t-SNE creates the low-dimensional embedding from the high-dimensional embedding by minimizing the Kullback-Leibler (KL) divergence of the joint probabilities in the original space and the embedded space using gradient descent[2].

t-SNE is chosen for further dimensional reduction for visualization purposes because it is highly sensitive to local structures. In other words, entities that are closer together in their high-dimensional representation are also placed closer together by t-SNE in the lower dimensional space. This is a desirable feature because we want to investigate whether news sources that close together in their high-dimensional representations are in fact similar. We can investigate this by looking at how news sources are placed in the lower 3-dimensional space and visually examine if news sources that are closer together are in fact similar.

t-SNE has a lot of parameters that can be tuned as an effort to find a better fit. Figure X below shows the complete parameters used in the t-SNE model used to visualize the news source representations.

[1]https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE
[2]https://scikit-learn.org/stable/modules/manifold.html#t-sne

```
{'angle': 0.5,
 'early_exaggeration': 12.0,
 'init': 'warn',
 'learning_rate': 'auto',
 'method': 'barnes_hut',
 'metric': 'euclidean',
 'min_grad_norm': 1e-07,
 'n_components': 3,
 'n_iter': 1000,
 'n_iter_without_progress': 300,
 'n_jobs': None,
 'perplexity': 50,
 'random_state': None,
 'square_distances': 'legacy',
 'verbose': 0}
```

Figure 7-1: The complete t-SNE parameters used in this thesis to visualize the news sources representations by projecting them down to three dimensions.

Using visual examination, I find that news representation as described in this thesis does indeed place 'similar' news sources closer together. However, the notion 'similar' is not constant throughout the distribution of the news sources in the representational space. For instance, some news sources that are closer together are similar based on their reliability (i.e. some unreliable news sources placed closer together). Yet, not all of the unreliable news sources are close together. Some of them are closer to other news sources instead, on the basis of similar country of origin, theme, media type, or other characteristics.

Some groupings of similar news sources are as follows:

1. thefederalist.com, redstate.com, pjmedia.com, breitbart.com, oann.com, thenationalpulse.com, americanthinker.com, amgreatness.com, bizpacreview.com, independentsentinel.com, theepochtimes.com are all unreliable news sources that spread misinformation and fake news, more recently about Covid-19 and the election. All of these news sources are close together in the representational space. The cluster containing these news sources are shown in Figure 7-2.

2. Canada-based news sources such as nationalpost.com, cbc.ca, winnipeg.ctvnews.ca, macleans.ca, winnipegfreepress.com, calgaryherald.com, torontosun.com, wind-

sorstar.com are close together in the representational space. In this case, the 'similarity' criterion of these news sources are their country of origin. This is as opposed to media type as they have different media types of online newspapers and TV network, or political leaning of conservative (calgaryherald.com, torontosun.com), liberal (macleans.ca), and unbiased (winnipegfreepress.com).

3. Some left-leaning news sources such as forward.com, thewalrus.ca, harpers.org, lamag.com, washingtonian.com, nhregister.com, and ctpost.com are close together despite their different media types of magazines (lamag.com, washingtonian.com), daily newspapers (ctpost.com, nhregister.com); country of origin of Canada (thewalrus.ca) and the USA; themes of culture, food, fashion (lamag.com), and Jewish arts culture and opinion (forward.com). In this case, the similarity criterion is political leaning.

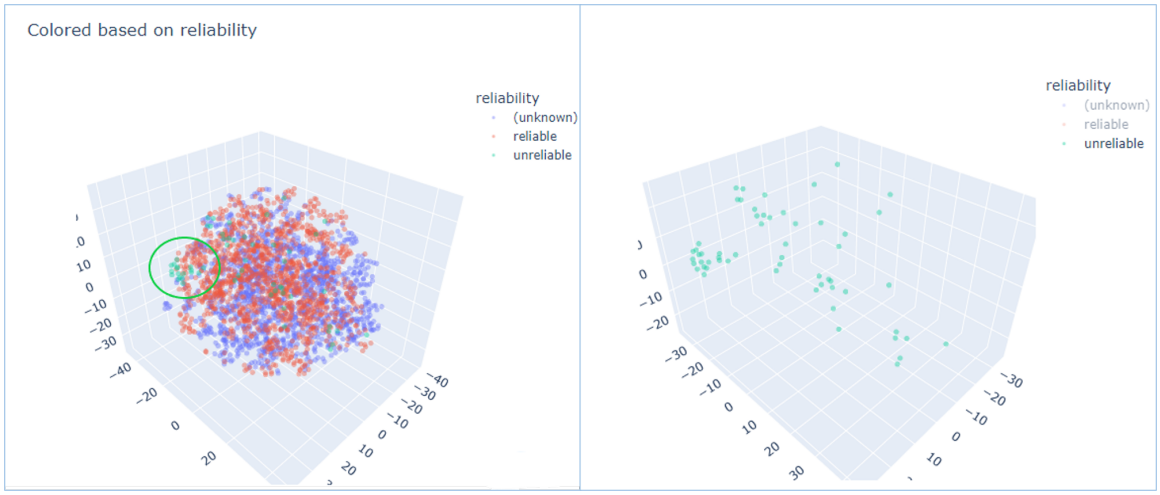

Figure 7-2: News source representations in 3D with colors representing their reliability (green for unreliable, red for reliable, and purple for unlabeled news sources). In the left side figure we see that news sources are not exactly grouped together based on their reliability. However, some unreliable news sources (enclosed by the green circle) are very close together. The same view with only unreliable news sources is presented by the figure on the right for clarity. Interactive three dimensional representations accessible at `https://newssource-vis.herokuapp.com/`.

Other groupings can be found by visual examinations of the interactive 3D representation of these news sources on https://newssource-vis.herokuapp.com/. The

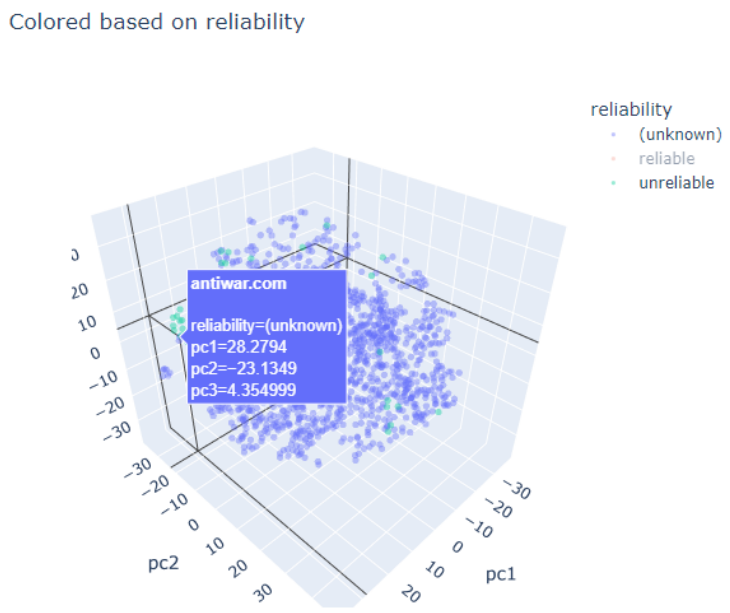

Figure 7-3: 'antiwar.com' is a close neighbor of a group of unreliable news sources. An intuitive suggestion is that perhaps 'antiwar.com' is an unreliable source. After manual investigation and according to Media Bias Fact Check as well as Politifact, 'antiwar.com' indeed has medium to low reliability.

resulting visualization of news sources' representations is shown in Figure 7-2, with colors corresponding to the reliability of the news sources as labeled by Media Bias Fact Check[3].

One potential application of this representation method is to help characterize unlabelled news sources by leveraging known facts about the neighboring news sources. For example, 'antiwar.com' is one of the unlabeled news sources in terms of their reliability. As shown in Figure 7-3, 'antiwar.com' is a close neighbor of the group of unreliable news sources as listed in number 1 in the list above. One can reasonably argue that since 'antiwar.com' has similar representations based on audience-based metrics as a group of unreliable sources, perhaps 'antiwar.com' is another unreliable source. After manual investigation and according to fact-checking websites such as Media Bias Fact Check as well as Politifact[4], 'antiwar.com' indeed has medium to low reliability. Perhaps other characterization tasks for unlabeled news sources could be done with this method.

[3]https://mediabiasfactcheck.com/
[4]https://www.politifact.com/

7.2 Clustering

After seeing that some similar news sources are indeed close together in the new feature space, in this section I investigate whether there are naturally emerging clusters of news sources based on the new feature space. The idea is to see what characteristics of news sources are the most prominent in clusters of news sources, if any, when these news sources are represented solely by their audience-based metrics as features.

In the following sections, I will give an overview of the two clustering methods that are considered in this thesis, and the silhouette score metric as a way to evaluate them. Next, I will expand on my attempt to cluster first using manually-chosen specific subreddits (utilizing agglomerative clustering), then using the entire set of subreddits (utilizing k-means clustering), to see if news source clusters emerge.

7.2.1 Clustering Algorithms Overview

There are various types of clustering algorithms, such as density based clustering (e.g. DBSCAN, HDBSCAN, etc.), hierarchical clustering (Agglomerative or Divisive), centroid-based partitioning clustering (K-Means), fuzzy clustering, and others. The first three types of clustering are hard-clustering, which means that each point is only assigned to one cluster, whereas fuzzy clustering is a soft clustering method that assigns probabilities to each point of joining the various clusters. In this thesis I focus on the hard clustering methods, specifically agglomerative clustering and k-means clustering.

For many clustering tasks, including the one that is being considered in this thesis, one does not know in advance to which cluster each of one's data points belong. In other words, the ground truth labels of the data points are not known. This means that it is hard to evaluate clustering results because there are no ground truth labels to compare the results to. One possible evaluation metric in this case is called the 'Silhouette Coefficient' metric as proposed by Rousseeuw in 1987 [36], which measures how well samples are clustered with other samples that are similar to themselves.

The Silhouette Coefficient s for one data point is given as $s = \frac{b-a}{max(a,b)}$, where a is

the mean distance between that data point and all the other points in the same class, and b is the mean distance between that data point and and all the other points in the next nearest cluster[5]. Intuitively, a measures cluster cohesion (where smaller a means better cohesion), while b measures cluster separation (where larger b means better separation). Thus better clustering results would lead to larger numerator and larger silhouette coefficient overall. Silhouette Coefficient values are in the range of -1 and 1, where higher values closer to 1 indicate better clustering. Values closer to -1 indicate assignments to the wrong cluster and values closer to 0 indicate overlapping clusters. One can calculate and report the Silhouette Coefficient value for each of the data points individually, or as one number representing the mean of all the Silhouette Coefficient values.

Next I will give an overview of the two clustering methods I consider in this thesis: agglomerative clustering and k-means clustering.

Agglomerative Clustering

Agglomerative clustering is a bottom-up hierarchical clustering approach, where initially each of the points is clustered in a 1-element cluster, then clusters are combined together at each step based on a particular distance and linkage metrics until eventually we end up with one big cluster containing all of the points.

When considering which clusters to merge in the intermediary agglomerative clustering method, the algorithm measures the distance between clusters and merge two closest clusters together at a time. Some common distance metrics are Euclidean distance, squared Euclidean distance, Manhattan distance, Chebyshev (maximum distance), and others. The usual default of distance metric is the Euclidean distance.

Since clusters generally contain multiple points (except for the initial clusters where each cluster only contains one point), measuring distance between two clusters is not very trivial. This is where the linkage criteria comes into play, where having different linkage criteria means measuring the distance between two clusters differently.

[5]https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation

There are various linkage criteria in agglomerative clustering, including the most popular as follows:

1. Single (minimum) linkage clustering considers the minimum distance between all possible pairings of points in the two clusters as the distance between the two clusters.

2. Complete (maximum) linkage clustering considers the maximum distance between all possible pairings of points in the two clusters as the distance between the two clusters.

3. Average linkage clustering considers the maximum distance between all possible pairings of points in the two clusters as the distance between the two clusters.

These linkage criteria are visualized in the Figure 7-4.

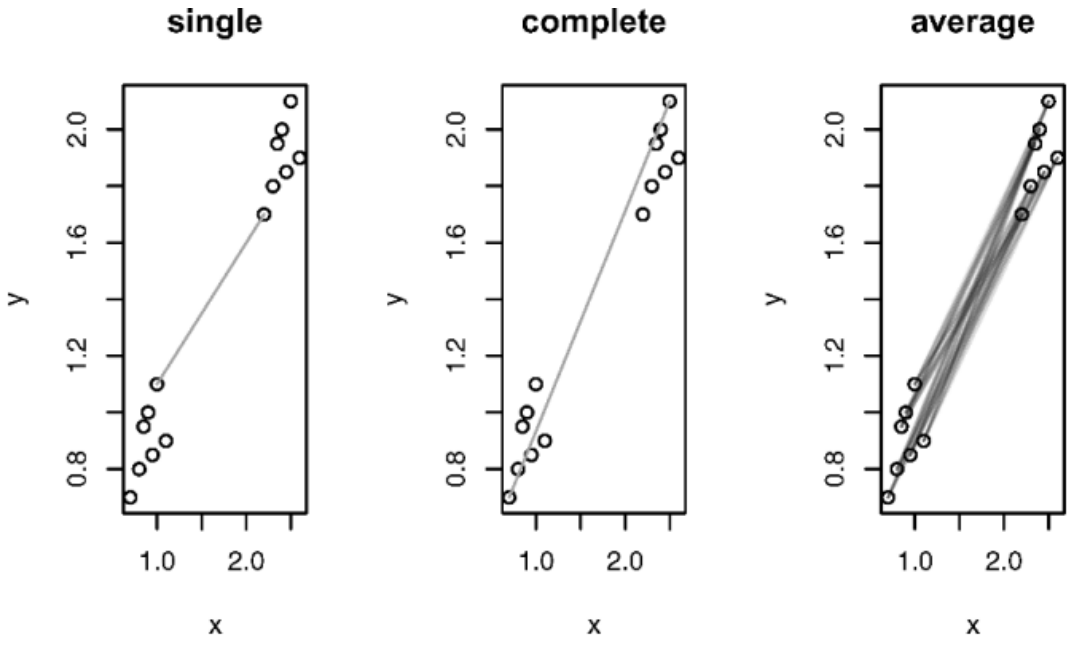

Figure 7-4: Different ways of calculating the distance between two clusters in agglomerative clustering, depending on the linkage criterion. Source: Figure 6.2 of Everitt et al. (2011) [15]

Some of the advantages of using hierarchical clustering such as agglomerative clustering are that the clustering process itself is easy to describe and visualize using a dendogram, and that it is non-parametric so we don't have to initially set

how many clusters we are trying to find, which is unlike another clustering method called k-means. However, agglomerative clustering algorithms are computationally expensive. Some variants such as SLINK [39] for single-linkage and CLINK [11] for complete-linkage have been developed to improve the runtime of agglomerative clustering algorithms[6].

K-Means Clustering

K-means clustering is an expectation-maximization algorithm aiming to cluster entities by minimizing the within-group sum of squared distances (WGSS) of the points in each cluster. Specifically:

$$WGSS = \sum_{j=1}^{q} \sum_{l=1}^{k} \sum_{i \in G_l} (x_{ij} - \bar{x}_j^{(l)})^2, \tag{7.1}$$

where $\bar{x}_j^{(l)} = \frac{1}{n} \sum_{i \in G_l} x_{ij}$ is the mean of the data points in group G_l on variable j [15].

The k-means clustering algorithm can be described as follows:

1. Initialize the centroids of the pre-determined k clusters. This can be done by randomly sampling k of the data points, or more developed initialization algorithms such as the k-means++ initialization that is used to speed up convergence[7].

2. Repeat until centroids do not change significantly or at all (convergence):

   (a) Assign each of the data points to the nearest centroid to it, forming k clusters.

   (b) For each of the k clusters, calculate its new centroid by taking the mean of all the data points.

Due to the randomness of the initialization process, different runs of k-means clustering algorithms could lead to different clustering results. Another challenge

[6] https://en.wikipedia.org/wiki/Hierarchical_clustering
[7] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.kmeans_plusplus.html

in using k-means clustering is the fact that we must predetermine the number of clusters k that we expect, as a parameter to the algorithm. In unsupervised cases where clustering algorithms are used, such as the case in this thesis, one does not know in advance how many clusters to expect, thus pre-determining k is not trivial. One way to do so is by employing the so-called 'elbow method' heuristic, where one plots the WGSS associated with different cluster sizes, then picking the elbow of the curve, where adding one more cluster does not result in significant WGSS decrease, as the optimal number of clusters[8].

The k-means clustering algorithm is not without advantages. This algorithm is one of if not the most popular and easiest-to-explain-and implement clustering algorithms with applications in a vast range of fields[9]. Unlike agglomerative clustering, k-means also scales well to large datasets, like the one in this thesis.

Now that we are more familiar with the agglomerative and k-means clustering algorithms, in the next sections I will go over my attempts to cluster the gm_intersection news sources to see if any of these news sources form natural clusters using their sharing metrics on Reddit as their features. I first realize clustering for a smaller and more explainable subset of my data, then for the whole dataset.

7.2.2 Small Scale: Clustering using sports subreddits

As the first step in clustering, I investigate whether using specific subreddits as features could yield insightful clustering results. Specifically, I use three types of subreddits: 33 NFL-related subreddits, 30 NBA-related subreddits, and 20 Premier League-related subreddits, as listed by Solomon[10]. In total, there are 85 news sources that are mentioned at least 5 times in one or more of these subreddits.

Short overview of the leagues:

- The National Basketball Association (NBA) is a professional basketball league in North America founded in New York in 1946[11], thus the subreddits related

[8]https://en.wikipedia.org/wiki/Elbow_method_(clustering)
[9]https://en.wikipedia.org/wiki/K-means_clustering
[10]https://www.futuressport.com/en/insight/sports-fans-on-reddit.aspx
[11]https://en.wikipedia.org/wiki/National_Basketball_Association

to this league are those that are specific to various regional basketball teams in North America, such as r/bostonceltics, r/NYKnicks, r/torontoraptors, and others.

- The National Football League (NFL) is a professional American football league founded in Ohio, USA in 1920[12]. Related subreddits are football teams such as r/Patriots (for Boston-based Patriots team), r/GreenBayPackers (for Wisconsin-based Green Bay Packers), r/Seahawks (for Seattle Seahawks), and others.

- The Premier League is the top level of the English football league system[13]. Some subreddits related to this league include r/reddevils (for Manchester United), r/LiverpoolFC, r/chelseafc, and others.

Using only these particular subreddits, I built representations of the 85 news sources using the method described in the Building News Sources Representation section. After running the PCA step, I found that the first 7 principal components explain 90% of the feature variance. So for clustering, I only use these 7 principal components.

For this clustering task, I first utilize the agglomerative clustering as implemented by the SciPy package in Python[14], with 'average' linkage. I then build a dendrogram based on this agglomerative clustering to visualize how news sources are clustered together in the intermediary steps of the algorithm. I then also realize the k-means algorithm for the same task and investigate whether either or both of the clustering algorithms yields interesting results.

From Figure 7-5 we see that the news sources are reasonably clustered based on their country of origin and themes (type of sports and whether the news sources are exclusively talking about sports or more general). Specifically, we see that there are 4 main clusters that show up: general news sources that sometimes writes about sports (represented by orange branches), UK-based news sources (represented by

[12]https://en.wikipedia.org/wiki/National_Football_League
[13]https://en.wikipedia.org/wiki/Premier_League
[14]https://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html

green branches), sports-specific or general news sources that heavily writes about sports (represented by the red branch), and nba.com in its own cluster.

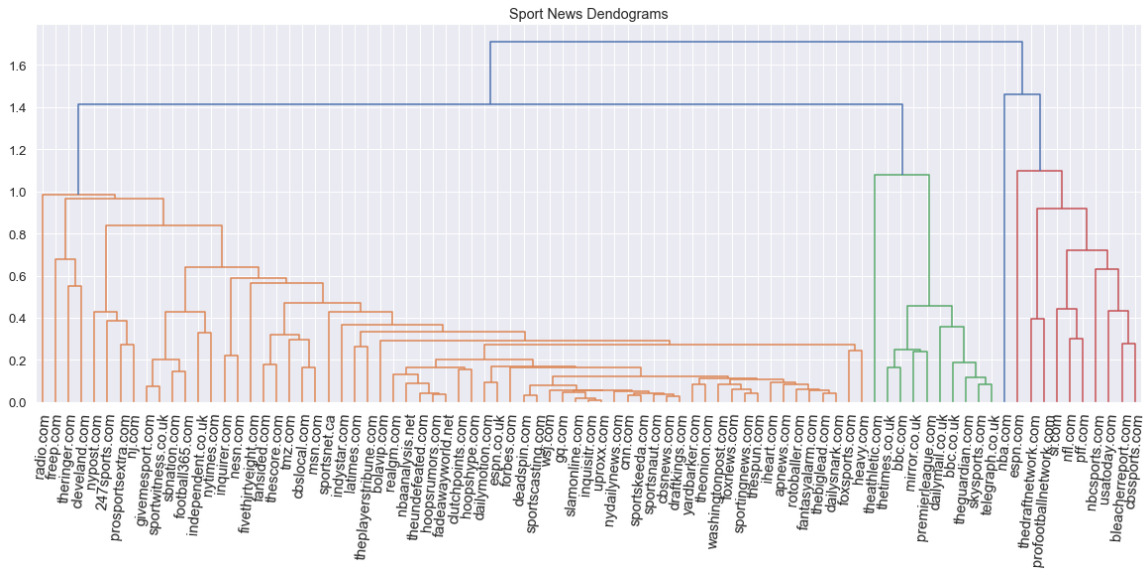

Figure 7-5: Dendrogram of the resulting agglomerative clustering of news sources that are shared in sports-related subreddits (NFL, NBA, Premier League), with weighted-linkage. We see that most UK-based news sources are clustered together (green branches), news sources that heavily write about basketball and American football are clustered together (red branches), and the more general news sources clustered together (orange). However this is not a perfect clustering since some UK-based news sources as well as sports-themed news sources are placed in the orange more general (orange) cluster.

The next step is to realize k-means clustering. For initialization, I use k-means++ initialization [42]. to speed up convergence. To find the optimal number of clusters, I employ the elbow method with results as shown in Figure 7-6. We see that the elbow of the WGSS curve is found for cluster size $k = 3$. Thus in the final cluster, I will realize k-means clustering on this small dataset using cluster size $k = 3$.

Similar to the result of agglomerative clustering, using k-means clustering of $k = 3$ results in one big cluster containing 63 general news sources, one cluster containing 11 UK-based news sources, and 11 news sources that writes heavily about sports.

As an evaluation metric for this clustering result, I inspect the silhouette coefficients of my 85 news source data points. The individual silhouette coefficient values are as shown in Figure 7-7, while the overall mean of the silhouette coefficient is 0.61.

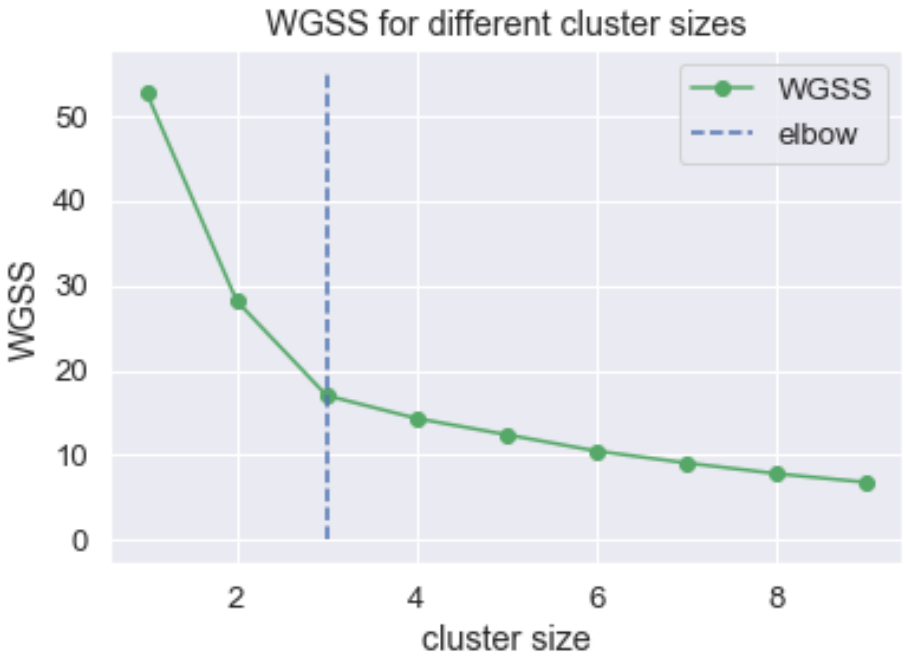

Figure 7-6: Different WGSS values for different cluster sizes for the small dataset acquired by only considering sports-related subreddits. We see that the elbow of the WGSS curve is found at cluster size k = 3, as after this point an additional cluster does not significantly decrease WGSS.

Recall that Silhouette Coefficient takes values between -1 and 1, so a value of 0.61 indicates a reasonably good clustering result.

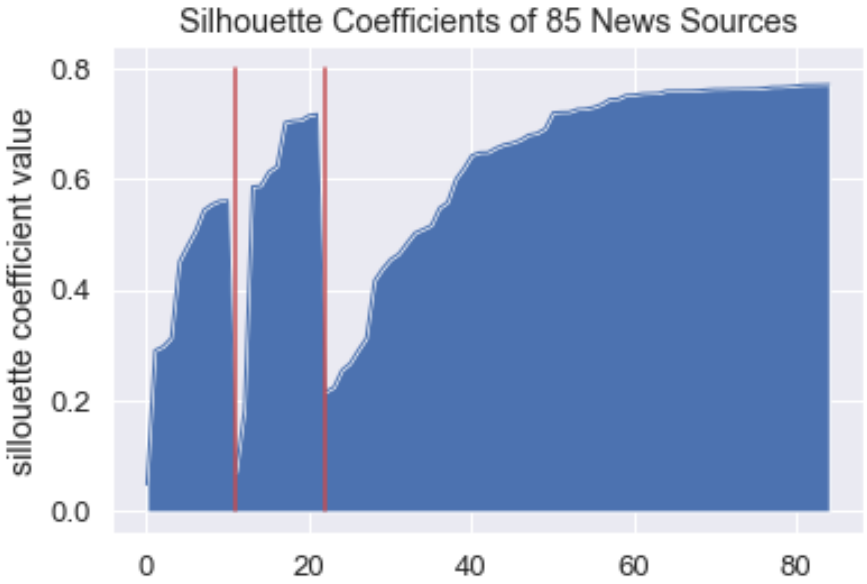

Figure 7-7: The silhouette coefficient values for individual data points. The red lines separate the three different clusters. The first cluster contains the news sources that heavily write about sports, the second cluster contains UK-based news sources, and the third cluster contains the remaining news sources that are general and only sometimes write about sports.

This clustering result is rather stable across different algorithms (k-means and agglomerative clustering). It is interesting to note that such clustering of news sources that differentiates theme and country of origin of news sources, emerges with merely news sources sharing statistics on Reddit as their features.

7.2.3 Big Scale: K-Means clustering for the entire dataset

We have seen that a reasonable clustering result of news sources emerges when we only use their sharing statistics on Reddit as their representational features, at least if we are intentional about choosing which subreddits to include, like we do in the Small Scale clustering section. In this section I investigate whether, when using all the available subreddits as features, any meaningful clustering results of news sources would emerge.

For this big scale clustering task, I utilize the k-means clustering algorithm, with k-means++ initialization. However, even after tuning the hyperparameters of the clustering algorithm, the clustering results do not seem very significant. First of all, in choosing the appropriate number of clusters k, the elbow method seems not to work well. As shown in the Figure 7-8, there is no discernible elbow of the curve to be found, as increasing the number of clusters from 2 to 100 results in a constant decrease of the WGSS. In other words, it seems that there is no one optimal number of clusters that can be found using this method.

Figure 7-8: Different WGSS values for different cluster sizes for the whole 2,647 news source dataset with all 2,189 subreddits used in building their representations. There is no obvious elbow of the WGSS curve.

As a second attempt to find the optimal number of clusters, I also calculated the mean silhouette score values for all numbers of clusters between 2 to 100. As shown in Figure 7-9 shows that the highest silhouette coefficient value is achieved with $k = 28$, with silhouette coefficient value being 0.2 which is a rather low value. Although, in general, these values are rather low in the scale of -1 and 1, I will investigate the clustering results for when $k = 28$.

Looking into the clustering results of the k-means algorithm with $k = 28$, some clusters indeed have meaningful prominent characteristics, while some others do not.

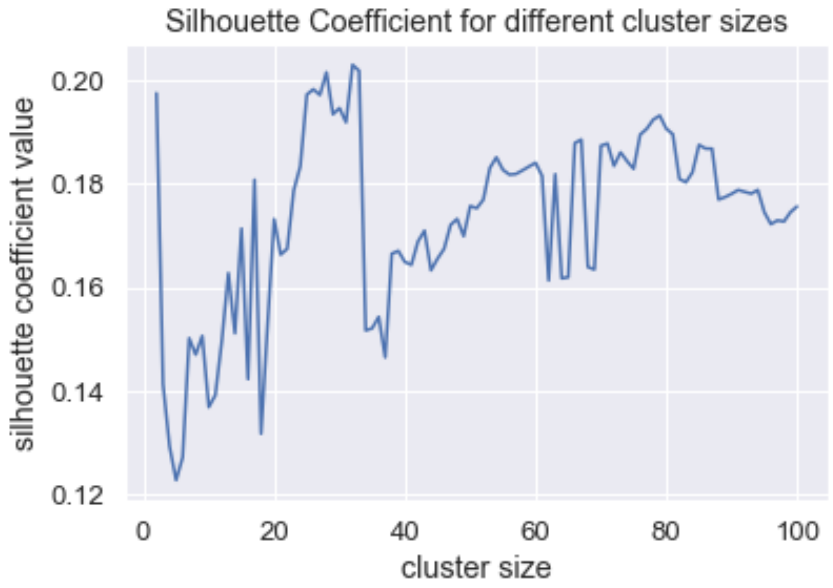

Figure 7-9: Silhouette Coefficient values for different cluster sizes. The highest silhouette coefficient value is achieved with k = 28, with silhouette coefficient value being 0.2. However, in general these values are rather low.

Similarly, some clusters have rather high silhouette coefficient means, while others have low means. Table A.1 in Appendix A describes prominent characteristics found in each cluster, if any, as well as silhouette coefficient means for the different 28 clusters, while Figure 7-10 presents the different silhouette coefficients of the members of each of the clusters. It is interesting to note how news sources specific to some regions, themes, or ideology come to be clustered together using Reddit audience-based metrics. The criterion that news sources in the same cluster share is different across clusters, and that the groups within that criterion do not all appear to have their own clusters. For example, there is one cluster of far-right news sources, but no cluster of far-left news sources. There is one cluster of TV-based news sources, but no cluster of podcasts.

Overall, based on their silhouette coefficient values, we see that some clusters are well defined while others are not. The low quality and interpretability of the clustering results could be due to either the lack of natural clustering of news sources using this representation method, or due to the need of further data processing before clustering.

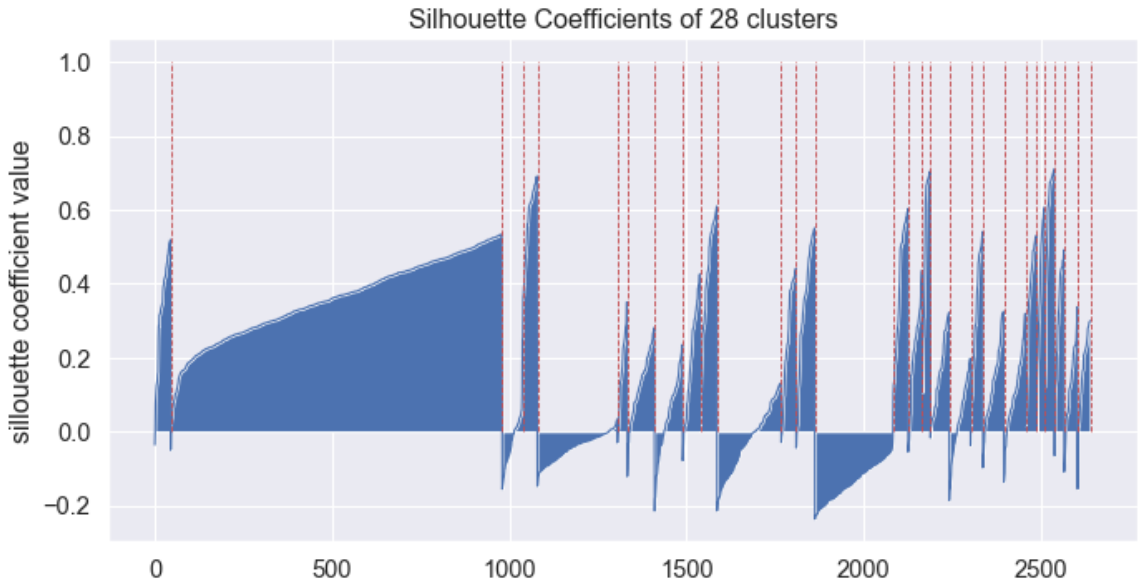

Figure 7-10: Silhouette coefficient values of news source data points in the 28 different clusters. The red lines separate the different clusters. Some clusters have reasonably good silhouette coefficients while some others have low and therefore bad silhouette coefficient values.

7.2.4 Clustering Conclusions and Challenges

Using the silhouette coefficient metric, we see that small-scale clustering using predetermined subreddits to build news source representations lead to better clustering quality compared to clustering results using all of the available subreddits. However, for both cases, the results are not perfectly interpretable, since, for example, in both cases the majority of the news sources being considered are clustered into one big cluster which indicates lack of natural clustering structure for most of the data points. Even so, some clusters with high silhouette coefficient values, indicate that some news sources are indeed close together in the space built based on their Reddit sharing statistics as their features. This could serve as a measure of news source similarity, where the notion of similarity takes the form of different characteristics such as country of origin, media type, political leaning, and others. Perhaps with more information incorporated into their representations, a better clustering of the news sources can be achieved. This information could include, for example, the sharing frequency of these news sources in the comment section, the text-based information

provided in the actual submissions and comments on Reddit as well as the subreddit descriptions.

7.3 Building Classifier: A Small Case Study

Another way to see whether we would be able to infer some specific characteristics of news sources (such as reliability, country of origin, political leaning, and media type) using the 300-dimensional representation as presented in this thesis, is by investigating whether, for a particular characteristic, we can build a model based on some labeled news sources to predict the label of the unlabelled ones. It turns out that a simple classification model can be built to predict particular characteristics, but not others. In this section I focus on two characteristics: country of origin and reliability.

The first characteristic that I focus on is country of origin. Using Wikidata[15], I labeled the news sources based on which country they are based in, and extracted 114 news sources that are based on one of the three countries Australia, France, and Germany. These three countries are chosen because based on the three dimensional representation of news sources built using t-SNE in the Visual Investigation section, they are reasonably well separated as seen in Figure 7-11, and they have a rather balanced number of news sources within each group (28 news sources from France, 35 news sources from Germany, and 51 news sources from Australia).

As a result, I was able to build a Support Vector Machine multiclass classifier that can classify these news sources based on their country of origin of three possibilities (Australia, France, and Germany) with 74% accuracy, with ROC AUC of 0.96. Although not extremely high, this accuracy is considerably higher than chance. The high ROC AUC (which computes the AUC of each class against the rest, also known as one-vs-rest, as implemented by sklearn[16] indicates that the model has a reasonably good predictive power. Note that I built this model using the first 20 principal components of the news source features, which cumulatively explain 40% of the feature

[15]https://www.wikidata.org/wiki/Wikidata:Main_Page
[16]https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html

variance. I found that using the first 20 principal components results in a better classification result, compared to if I further reduce the 300 principal components I have been using for analysis down to 20 features.

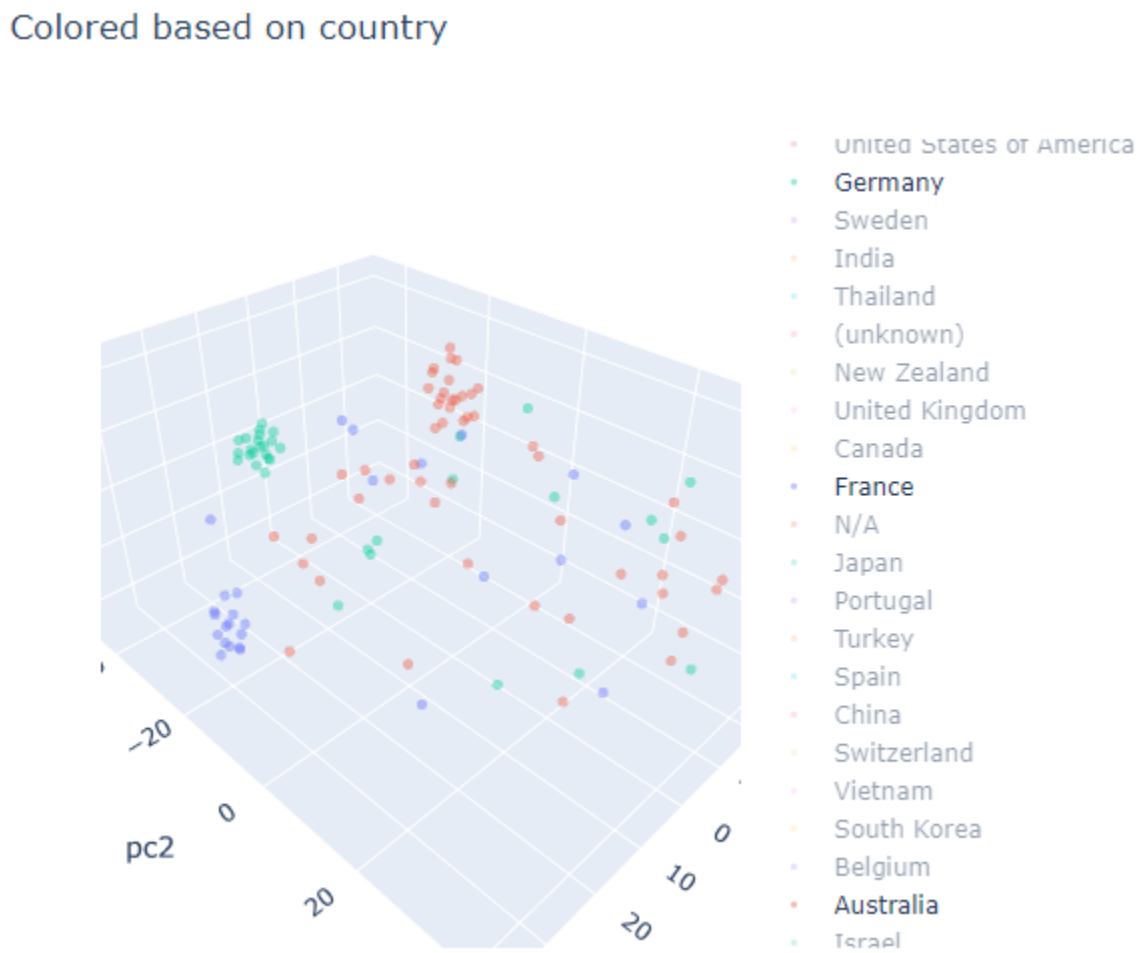

Figure 7-11: The three dimensional visualization of news sources based in Australia, France, and Germany seem to be well (although not perfectly) separated. The green dots represent Germany-based news sources, the purple dots represent France-based news sources, and the red dots represent Australia-based news sources.

However, for other characteristics such as reliability, building a predictive model is not as straightforward. With Media Bias Fact Check[17] as the source of credibility labels, the models that I have come up with have not been particularly useful in predicting whether news sources are reliable or not using the representation method outlined in this thesis. The various models I built (using various regression and tree

[17]https://mediabiasfactcheck.com/

based models) have very low accuracies of <=50%, which is no better than chance.

There are a couple possible reasons for the poor performing model for news source reliability prediction. The first one is that perhaps further feature engineering and more complex models are needed to use this type of news source representation to build a model to classify and predict news source reliability. The second possible reason is that it is simply hard if not impossible to infer the reliability of most news sources by only looking at their audience-based metrics. Previous research projects on news reliability/credibility have instead focused on the news article level [40, 19], instead of news source level as in this thesis. Different articles published by the same news source can have different reliability/credibility thus it might make more sense to infer this quality in a more granular news article level.

At least based on this small case study, it is interesting to note how particular characteristics such as country of origin (for specific countries) is much easier to predict than other characteristics such as reliability, using the representational method built based on Reddit audience sharing metrics. Further studies could look into the goodness of models that can be built for the multiclass classification task including all the possible countries, and classification tasks for other characteristics such as political leaning or media type.

Chapter 8

Conclusion and Future Work

8.1 Conclusion

By visualizing the representations of the gm_intersection news sources built using their Reddit audience-based metrics, I find that the news sources that are close together in the representational space indeed have some "similar" characteristics, although the notion of similar is different from one neighborhood of news sources to another. Some news sources that are closer together are based in the same countries, while others that are closer together have a shared theme, yet others are closer together with shared political leanings or reliability. Therefore, for unlabeled news sources, one can potentially observe where they are located in the representational space built with the Reddit audience-based metrics as outlined in this thesis, and infer particular characteristics of those unlabeled news sources by looking at the neighboring news sources.

In addition to visual examination of the representations of the news sources, I also attempt to cluster them together using their Reddit audience-based sharing statistics. Although imperfect, the clustering results suggest that particular news source characteristics such as country of origin and some specific themes are prominent features that give rise to cluster structures that are easier to identify using k-means algorithms, than other features. Similarly, a small case study to build a simple classifier suggests that building a classifier for predictive tasks is easier accomplished for some features

than others.

In conclusion, there is some evidence that Reddit audience sharing statistics alone can be used in inferring some characteristics of news sources. Even a simple exploration of comparing the total sharing frequency of a news source and the number of subreddits mentioning those news sources leads to interesting insights about news source popularity and the broadness of their audience, as shown in the Exploring News Sources section. However, more work needs to be done to further engineer or improve these features and potentially build models to predict these characteristics for unlabelled news sources.

8.2 Future Work

In this section, I will go over things that can be improved on and various methods that can be implemented to further the analysis in this thesis.

First of all, I excluded many news sources because they were infrequently shared in the dataset I utilized. Using more Reddit data, instead of only from January to June 2021 as in this thesis, could provide more data points for these news sources, allowing them to be included meaningfully in the analysis. Further analysis should also be done to investigate what kinds of news sources are shared so infrequently on Reddit and if they are different from the more popular news sources in any way. Moreover, through the data exploration, I found that there are a lot of blogs and podcasts being considered as news sources by both GDELT and Muck Rack. Further studies should investigate the relevance of these user-generated content hosts in the online news ecosystem.

Second, more information from Reddit could be included in building audience-based news source representations. For example, recall that each Reddit user has their own Karma score, which is potentially useful information. This thesis aggregates user engagement with news sources by only looking at specific subsets of users. Future work can directly measure audience engagement by considering different users separately. Additionally, according to Glenski et al., most people do not read an

article prior to voting [17] and according to Morrison, Reddit users are not all equally likely to vote on submissions and comments [28], thus, a different way of incorporating voting information could be considered to take these findings into account. Moreover, this thesis also only considers the submissions on Reddit that contain news source links. Future work could also consider comments that contain news source links (instead of just counting comments per post without investigating the content of those comments). Last but not least, this thesis does not make use of various text-based information from Reddit, such as the actual submission texts, user profile descriptions, and subreddit descriptions. Various NLP techniques could be used to incorporate these.

Third, in terms of analysis, several other methods should be employed for further analysis. For example, for further processing news source features, including to reduce the feature space dimensions, more complex methods could be used. In this thesis I mainly use PCA, which is a linear dimensionality reduction technique, however non-linear dimensionality reduction techniques should also be considered. In this thesis, I use one such non-linear dimensionality reduction technique, which is t-SNE. However, unlike PCA, t-SNE is intended solely for visualization, not for data preprocessing [41]. Another non-linear dimensionality reduction technique worth considering is an autoencoder, which is a type of neural network (with non-linear activation functions) that aims to meaningfully encode input vectors into a lower dimensional latent space, and decode it back to a reconstructed input such that it is as similar as possible as the original input [3]. Mean Squared Error (MSE) is the common loss criterion used to train autoencoders. The lower the loss, the more similar the input and reconstructed input are. In fact, I have attempted to implement one myself for this thesis, using the pytorch framework, which is an open source machine learning framework [1], resulting in an autoencoder with low reconstruction MSE error (0.0013). However, the resulting latent representations did not yield useful results in the clustering or classification tasks, thus I conclude that more work needs to be done to build a useful auto-encoder model for this task.

[1] https://pytorch.org/

Furthermore, for clustering tasks, other clustering algorithms could be used to detect emerging news source cluster structure. In particular, I have only considered hard-clustering which assigns each point to exactly one cluster, yet soft-clustering such as fuzzy clustering [2] that assigns probabilities to points belonging to each of the different clusters could also be considered. In terms of k-means clustering itself, a variant called weighted k-means that specifically caters to sparse data [21] could also be considered. Yet another clustering algorithm that should be considered in future works is model-based clustering. Landau et al. point out that even though there are not much objections to the use of k-means or agglomerative clustering, model-based clustering is the one that would give most persuasive results (at least to statisticians) in terms of formal inference [24].

Lastly, a method of analysis that should be utilized is analysis using network science. Recall that I represent news sources using their sharing statistics in Reddit in a data frame or matrix where each row records one news source while one or multiple columns record the news source's sharing statistic in one subreddit. We can frame this problem using a graph data structure where the nodes represent news sources, and the edges between news sources represent their connection in terms of being shared in the same subreddit(s). The weights of the edges would depend on the number of common subreddits a pair of news sources are shared in, and the similarity of their sharing frequencies. The size of the nodes could be used to represent the popularity of news sources across subreddits. There are two ways of building such a graph: by directly building the graph as described above with a manually determined edge weight, or by first building a bipartite graph and projecting it down to such a graph as described above. In the case of the bipartite graph, there would be two types of nodes: one representing news sources, and the other representing subreddits. Edges would only exist between the two different types of nodes and not among the same type. This bipartite graph could then be projected down to the final graph as described above.

In fact, I have spent some time going in the direction of network analysis in this

[2] https://pythonhosted.org/scikit-fuzzy/auto_examples/plot_cmeans.html

thesis, although eventually not pursuing it further due to limited time. Figure 8-1, for example, shows a preliminary graph I built using only the top 200 subreddits ranked by the number of their subscribers. The size of the nodes is scaled to represent the number of total mentions of the associated news sources (the more a news source is shared then the bigger the node used to represent it), while the width of the edges represents the strength of connection or similarity between two news sources calculated using the number of common subreddits they appear together and the similarity of their sharing frequencies.

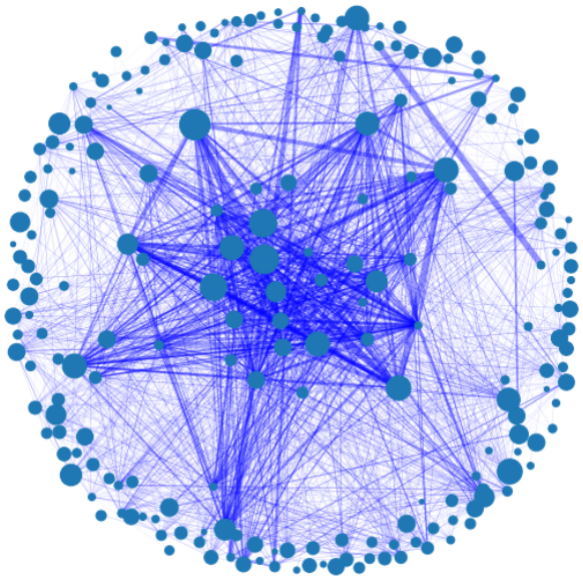

Figure 8-1: A preliminary graph built using only the top 200 subreddits ranked by the number of their subscribers. The size of the nodes is scaled to represent the number of total mentions of the associated news sources , while the width of the edges represents the strength of connection or similarity between two news sources calculated using the number of common subreddits they appear together and the similarity of their sharing frequencies.

Using such a representation, one can gain insights about the news sources that are similar to each other by investigating the ego-centric graph of each of the nodes, the popularity of news sources by investigating the size of the nodes and their degrees, and other insights. This is a promising direction for future research.

Appendix A

Big Scale Clustering Result

Cluster	Prominent Characteristics	Silhouette Coefficient	%news source in the cluster
1	Canada-based news sources (e.g. torontosun.com, toronto.ctvnews.ca, citynews.ca, calgaryherald.com, cbc.ca, etc.)	0.33	1.7%
2	-	0.36	35.4%
3	-	-0.03	2.1%
4	India-based news sources (e.g. theprint.in, businesstoday.in, thewire.in, thehindu.com, nationalheraldindia.com, etc.)	0.56	1.6%
5	-	-0.04	8.6%
6	News sources from various countries in Europe (e.g. greekreporter.com, thefirstnews.com, notesfrompoland.com, lrt.lt, politico.eu, intellinews.com, etc.)	0.17	1.1%
7	-	0.12	2.9%
8	Technology-related news sources (e.g. pcworld.com, arstechnica.com, geekwire.com, techspot.com, acm.org, etc.)	0.04	2.9%

Cluster	Prominent Characteristics	Silhouette Coefficient	%news source in the cluster
9	Comics, fantasy, and movies related news sources (e.g. variety.com, tor.com, looper.com, wegotthiscovered.com, ew.com, etc.)	0.19	1.9%
10	Game related news sources (e.g. nintendoeverything.com, kotaku.com, gematsu.com, altchar.com, dsogaming.com, etc.)	0.41	1.8%
11	TV-stations (e.g. wgrz.com, 14news.com, tmj4.com, wane.com, cheknews.ca, etc.)	-0.02	6.9%
12	Cars and clean energy related news sources (e.g. teslarati.com, rechargenews.com, greencarcongress.com, caranddriver.com, electrive.com)	0.26	1.6%
13	Far-right and unreliable news sources (e.g. redstate.com, theamericanconservative.com, theepochtimes.com, oann.com, lifesitenews.com, etc.)	0.33	2.0%
14	-	-0.14	8.3%
15	Cryptocurrency and fintech related news sources (e.g. cryptonews.com, blockonomi.com, ethereumworldnews.com, cointelegraph.com, coinspeaker.com, etc.)	0.40	1.6%
16	Stock market and investment related news sources (e.g. nasdaq.com, investorplace.com, morningstar.com, simplywall.st, benzinga.com, etc.)	0.21	1.5%
17	France-based and French-languages news sources (e.g. lemonde.fr, lefigaro.fr, journaldemontreal.com, arte.tv, radio-canada.ca, etc.)	0.53	0.9%

Cluster	Prominent Characteristics	Silhouette Coefficient	%news source in the cluster
18	Academia-related sites and publications (e.g. nature.com, psychologytoday.com, wiley.com, sciencedirect.com, bmj.com, etc.)	0.16	2.0%
19	Asia-based news sources (e.g. yna.co.kr, nhk.or.jp, taiwannews.com.tw, chinadaily.com.cn, coconuts.co, etc.)	0.03	2.2%
20	Texas-based news sources (e.g. houstonchronicle.com, cbsaustin.com, culturemap.com, dallasobserver.com, dmagazine.com, etc.)	0.32	1.3%
21	Science-related news sources especially magazines (e.g. popularmechanics.com, inverse.com, ecowatch.com, discovermagazine.com, aeon.co, etc.)	0.13	2.2%
22	News sources based in or about the middle east region and Russia (e.g. aljazeera.com, israelhayom.com, al-monitor.com, palestinechronicle.com, rt.com, etc.)	0.10	2.3%
23	Florida-based news sources (e.g. nbcmiami.com, miamiherald.com, floridapolitics.com, orlandoweekly.com, sun-sentinel.com, etc.)	0.37	1.2%
24	Massachusetts (mainly Boston) based news sources (e.g. nbcboston.com, bostonglobe.com, wbur.org, masslive.com, capecodtimes.com, etc.)	0.41	0.8%

Cluster	Prominent Characteristics	Silhouette Coefficient	%news source in the cluster
25	Germany-based news sources (e.g. spiegel.de, zeit.de, berliner-zeitung.de, tagesspiegel.de, focus.de, etc.)	0.57	1.1%
26	Defense and military related news sources (e.g. armytimes.com, defensenews.com, taskandpurpose.com, airforcemag.com, stripes.com, etc.)	0.31	1.0%
27	United Kingdom based news sources (e.g. telegraph.co.uk, theathletic.com, bbc.co.uk, mirror.co.uk, belfasttelegraph.co.uk, etc.)	0.12	1.4%
28	-	0.17	1.4%

Table A.1: Prominent characteristics found in each cluster, if any, as well as silhouette coefficient means for the different 28 clusters found using k-means clustering of the whole dataset.

Appendix B

Manually Removed Sites

amazon.com	quora.com	home.blog
wordpress.com	chase.com	archive.org
amazonaws.com	wattpad.com	herokuapp.com
bit.ly	t.co	mystrikingly.com
imgur.com	gofundme.com	xda-developers.com
soundcloud.com	paypal.com	redbubble.com
facebook.com	telegra.ph	artstation.com
spotify.com	googleapis.com	teespring.com
vimeo.com	web.app	notion.so
goodreads.com	octopus.energy	substack.com
goo.gl	shopify.com	over-blog.com
youtube.com	playstation.com	cloudfront.net
t.me	qualtrics.com	steamcommunity.com
google.com	naver.com	bandcamp.com
twitter.com	linkedin.com	kicksonfire.com
discord.com	samsung.com	discogs.com
pornhub.com	walmart.com	xvideos.com
amazon.co.uk	netlify.app	fangraphs.com
dropbox.com	android.com	coingecko.com

tumblr.com	wixsite.com	podbean.com
etsy.com	ikea.com	trello.com
github.com	eventbrite.com	bravesites.com
medium.com	yahoo.com	exercism.io
ebay.com	weworkremotely.com	bonhams.com
imdb.com	myspace.com	food52.com
blogspot.com	instagram.com	bookmyshow.com
pinterest.com	myshopify.com	withgoogle.com
patreon.com	whatsapp.com	nucypher.com
microsoft.com	anchor.fm	cpacanada.ca
deviantart.com	nintendo.com	hermanmiller.com
tiktok.com	homedepot.com	haproxy.com
netflix.com	audible.com	pixelstech.net
towardsdatascience.com	asus.com	line.me
squarespace.com	merriam-webster.com	tripsavvy.com
googleusercontent.com	intel.com	trivago.com
webmd.com	leagueoflegends.com	wikipedia.org
yahoo.com	nike.com	wikimedia.org
adobe.com	discordapp.com	wikiquote.org
apple.com	blogger.com	wikidot.com
weebly.com	office.com	wikisource.org
yahoo.co.jp	itch.io	cloudinary.com
weibo.com	lego.com	wikitionary.org
qq.com	mailchi.mp	fandom.com
crunchyroll.com	go.com	
amazon.de	snapchat.com	

Table B.1: A list of 133 sites I manually remove from my initial set of news sources, because some of these sites are more appropriately categorized as user-generated content hosts, and some others are not news related sites.

Appendix C

Example of One NDJSON Representation of Reddit Submission

The following is an example of a NDJSON representation of a Reddit submission.


```
{
    'all_awardings': [],
    'allow_live_comments': False,
    'archived': False,
    'author': 'elanglohablante9805',
    'author_created_utc': 1609519842,
    'author_flair_background_color': '#ffb000',
    'author_flair_css_class': None,
    'author_flair_richtext': [],
    'author_flair_template_id': '4f908eaa-9664-11ea-
                                a567-0ed46a42aec3',
    'author_flair_text': 'Historiador  80-Day Streak',
    'author_flair_text_color': 'dark',
    'author_flair_type': 'text',
    'author_fullname': 't2_9lr431i4',
    'author_patreon_flair': False,
```

```
'author_premium': False,
'can_gild': True,
'category': None,
'content_categories': None,
'contest_mode': False,
'created_utc': 1617235201,
'discussion_type': None,
'distinguished': None,
'domain': 'self.WriteStreakES',
'edited': False,
'gilded': 0,
'gildings': {},
'hidden': False,
'hide_score': False,
'id': 'mhj2hj',
'is_created_from_ads_ui': False,
'is_crosspostable': True,
'is_meta': False,
'is_original_content': False,
'is_reddit_media_domain': False,
'is_robot_indexable': True,
'is_self': True,
'is_video': False,
'link_flair_background_color': '',
'link_flair_css_class': None,
'link_flair_richtext': [],
'link_flair_text': None,
'link_flair_text_color': 'dark',
'link_flair_type': 'text',
'locked': False,
```

```
'media': None,
'media_embed': {},
'media_only': False,
'name': 't3_mhj2hj',
'no_follow': True,
'num_comments': 2,
'num_crossposts': 0,
'over_18': False,
'parent_whitelist_status': None,
'permalink': '/r/WriteStreakES/comments/
             mhj2hj/streak_90_ha_llegado_la_primavera/',
'pinned': False,
'pwls': None,
'quarantine': False,
'removed_by_category': None,
'retrieved_utc': 1623447663,
'score': 1,
'secure_media': None,
'secure_media_embed': {},
'selftext': 'Los p jaros est n cantando, las hierbas
            verdes est n brotando, y tengo alergias.
            Esto es la temporada de las alergias.
            Estornudo cada ma ana cuando me despierto, y
            otra vez si voy afuera. Necesito tomar
            medicina cada d a, pero no funciona tan bien.
            \n\nPor fuera, las lomas son bonitas porque
            son verdes y los robles tienen hojas nuevas.
            Por el fin de semana, hago caminatas pero
            cuando regreso a casa, necesito ducharme para
            remover el polen.\n\nCuando me jubile, voy a
```

```
            viajar al desierto cada a o por toda la
            primavera. No me gustar a quedarme aqu .',
'send_replies': True,
'spoiler': False,
'stickied': False,
'subreddit': 'WriteStreakES',
'subreddit_id': 't5_2eamt5',
'subreddit_subscribers': 2205,
'subreddit_type': 'public',
'suggested_sort': None,
'thumbnail': 'self',
'thumbnail_height': None,
'thumbnail_width': None,
'title': 'Streak 90: Ha llegado la primavera',
'top_awarded_type': None,
'total_awards_received': 0,
'treatment_tags': [],
'upvote_ratio': 1.0,
'url': 'https://www.reddit.com/r/
        WriteStreakES/comments/mhj2hj/
        streak_90_ha_llegado_la_primavera/',
'whitelist_status': None, 'wls': None}
```

Appendix D

Code Repository

The code I wrote to complete this thesis can be found in this GitHub repository:

`https://github.com/jsirait/honor_thesis`

Bibliography

[1] Eytan Bakshy, Solomon Messing, and Lada A Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015.

[2] Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. Predicting factuality of reporting and bias of news media sources. *arXiv preprint arXiv:1810.01765*, 2018.

[3] Dor Bank, Noam Koenigstein, and Raja Giryes. Autoencoders. *arXiv preprint arXiv:2003.05991*, 2020.

[4] Joana M Barros, Paul Buitelaar, Jim Duggan, and Dietrich Rebholz-Schuhmann. Unsupervised classification of health content on reddit. In *Proceedings of the 9th International Conference on Digital Public Health*, pages 85–89, 2019.

[5] Daniel A Bowen, Julie O'Donnell, and Steven A Sumner. Increases in online posts about synthetic opioids preceding increases in synthetic opioid death rates: a retrospective observational study. *Journal of general internal medicine*, 34(12):2702–2704, 2019.

[6] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–25, 2018.

[7] Usman W Chohan. Counter-hegemonic finance: The gamestop short squeeze. *Available at SSRN 3775127*, 2021.

[8] Nicola Daniele Coniglio, Vitorocco Peragine, and Davide Vurchio. The geography of displacement, refugees' camps and social conflicts. 2022.

[9] Brent D Davis, Dawn Estes McKnight, Rumi Chunara, Daniel J Lizotte, and Alona Fyshe. Quantifying depressed social media during covid-19: Information retrieval with ml & nlp. 2020.

[10] Munmun De Choudhury and Sushovan De. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Eighth international AAAI conference on weblogs and social media*, 2014.

[11] Daniel Defays. An efficient algorithm for a complete link method. *The Computer Journal*, 20(4):364–366, 1977.

[12] Venkatesh Duppada. "attention" for detecting unreliable news in the information age. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[13] William H Dutton, Bianca Reisdorf, Elizabeth Dubois, and Grant Blank. Search and politics: The uses and impacts of search in britain, france, germany, italy, poland, spain, and the united states. 2017.

[14] William H Dutton, Bianca Christin Reisdorf, Elizabeth Dubois, and Grant Blank. Search and politics: A cross-national survey.(2017). 2017.

[15] Brian Everitt and Torsten Hothorn. *An introduction to applied multivariate analysis with R*. Springer Science & Business Media, 2011.

[16] Michael Fire and Carlos Guestrin. The rise and fall of network stars: Analyzing 2.5 million graphs to reveal how high-degree vertices emerge over time. *Information Processing & Management*, 57(2):102041, 2020.

[17] Maria Glenski, Corey Pennycuff, and Tim Weninger. Consumers and curators: Browsing and voting patterns on reddit. *IEEE Transactions on Computational Social Systems*, 4(4):196–206, 2017.

[18] Mauricio Gruppi, Benjamin D Horne, and Sibel Adali. An exploration of unreliable news classification in brazil and the us. *arXiv preprint arXiv:1806.02875*, 2018.

[19] Priyanka Harjule, Akshat Sharma, Sachin Chouhan, and Shashank Joshi. Reliability of news. In *2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE)*, pages 165–170. IEEE, 2020.

[20] J Craig Jenkins and Thomas V Maher. What should we do about source selection in event data? challenges, progress, and possible solutions. *International Journal of Sociology*, 46(1):42–57, 2016.

[21] Liping Jing, Michael K Ng, and Joshua Zhexue Huang. An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. *IEEE Transactions on knowledge and data engineering*, 19(8):1026–1041, 2007.

[22] Igor Kotenko, Yash Sharma, and Alexander Branitskiy. Predicting the mental state of the social network users based on the latent dirichlet allocation and fasttext. In *2021 11th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, volume 1, pages 191–195. IEEE, 2021.

[23] Navin Kumar, Isabel Corpus, Meher Hans, Nikhil Harle, Nan Yang, Curtis McDonald, Shinpei Nakamura Sakai, Kamila Janmohamed, Keyu Chen, Frederick L Altice, et al. Covid-19 vaccine perceptions in the initial phases of us vaccine rollout: An observational study on reddit. 2022.

[24] Sabine Landau, Morven Leese, Daniel Stahl, and Brian S Everitt. *Cluster analysis*, section 6.5, page 185. John Wiley & Sons, 2011.

[25] Kalev Leetaru and Philip A Schrodt. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, volume 2, pages 1–49. Citeseer, 2013.

[26] Yang Liu, Christopher Whitfield, Tianyang Zhang, Amanda Hauser, Taeyonn Reynolds, and Mohd Anwar. Monitoring covid-19 pandemic through the lens of social media using natural language processing and machine learning. *Health Information Science and Systems*, 9(1):1–16, 2021.

[27] Pin Luarn and Ai-Yun Hsieh. Speech or silence: the effect of user anonymity and member familiarity on the willingness to express opinions in virtual communities. *Online Information Review*, 2014.

[28] Donn Morrison and Conor Hayes. Here, have an upvote: Communication behaviour and karma on reddit. *INFORMATIK 2013–Informatik angepasst an Mensch, Organisation und Umwelt*, 2013.

[29] Narjisse Nejjari, Sara Lahlou, Oumaima Fadi, Karim Zkik, Mustapha Oudani, and Houda Benbrahim. Conflict spectrum: An empirical study of geopolitical cyber threats from a social network perspective. In *2021 Eighth International Conference on Social Network Analysis, Management and Security (SNAMS)*, pages 01–07. IEEE, 2021.

[30] Orestis Papakyriakopoulos, Juan Carlos Medina Serrano, and Simon Hegelich. The spread of covid-19 conspiracy theories on social media and the effect of content moderation. *The Harvard Kennedy School (HKS) Misinformation Review*, 18, 2020.

[31] Rachel Parks, Emily C Newsom, Joyce H Park, and Naomi Lawrence. Skincare addiction on reddit: dermatology enthusiasts talk skin. *Dermatologic Surgery*, 46(10):1372–1374, 2020.

[32] Gordon Pennycook and David G Rand. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, 116(7):2521–2526, 2019.

[33] Fengcai Qiao, Pei Li, Xin Zhang, Zhaoyun Ding, Jiajun Cheng, and Hui Wang. Predicting social unrest events with hidden markov models using gdelt. *Discrete Dynamics in Nature and Society*, 2017, 2017.

[34] Ronald E Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. Auditing partisan audience bias within google search. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–22, 2018.

[35] Stephen Roller, Sainbayar Sukhbaatar, Jason Weston, et al. Hash layers for large sparse models. *Advances in Neural Information Processing Systems*, 34, 2021.

[36] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[37] Mattia Samory, Vartan Kesiz Abnousi, and Tanushree Mitra. Characterizing the social media news sphere through user co-sharing practices. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 602–613, 2020.

[38] Vinay Setty and Erlend Rekve. Truth be told: Fake news detection using user reactions on reddit. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 3325–3328, 2020.

[39] Robin Sibson. Slink: an optimally efficient algorithm for the single-link cluster method. *The computer journal*, 16(1):30–34, 1973.

[40] Francesca Spezzano, Anu Shrestha, Jerry Alan Fails, and Brian W Stone. That's fake news! reliability of news when provided title, image, source bias & full article. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–19, 2021.

[41] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[42] Sergei Vassilvitskii and David Arthur. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035, 2006.

[43] Veniamin Veselovsky, Isaac Waller, and Ashton Anderson. Imagine all the people: Characterizing social music sharing on reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 739–750, 2021.

[44] Yuping Wang, Savvas Zannettou, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, and Gianluca Stringhini. A multi-platform analysis of political news discussion and sharing on web communities. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 1481–1492. IEEE, 2021.

[45] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Nicolas Kourtelris, Ilias Leontiadis, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. The web centipede: understanding how web communities influence each other through the lens of mainstream and alternative news sources. In *Proceedings of the 2017 internet measurement conference*, pages 405–417, 2017.