UNIVERSITÀ
DI PARMA

Department of Mathematical, Physical and Computer Sciences

Master's degree course in Computer Sciences

Analisi delle Echo Chambers negli Aggregatori di Notizie

Study of Echo Chambers in News Media Aggregators

Graduating:
Pier Luigi Trespidi

Dep. Coordinator:
Flavio Bertini

Supervisors:
Rajesh Sharma
Roshni Chakraborty

Academic Year 2022–2023

Tartu, Estonia. March 2023

To Flavio, Rajesh and Roshni.

Contents

List of Figures

List of Tables

Abstract

L'evoluzione del comportamento di consumo delle notizie, guidata dalla diffusione predominante dei social media e delle piattaforme di notizie online, ha rivoluzionato il modo in cui il singolo individuo accede alle informazioni. Con un significativo spostamento verso i canali digitali, gli utenti hanno ora accesso illimitato a un flusso continuo di notizie, non vincolato dalle limitazioni dei tradizionali giornali stampati. Questa situazione impone alle fonti dei media di comprendere le preferenze del loro pubblico e di adattarsi di conseguenza. Nel soddisfare le esigenze degli utenti, le fonti dei media di informazione spesso allineano i loro contenuti alle preferenze degli utenti, favorendo l'esposizione selettiva e dando potenzialmente origine a camere d'eco (*echo chambers*). Questi ambienti virtuali rafforzano le convinzioni degli utenti, esponendoli ripetutamente a informazioni che si allineano con i loro specifici interessi. La manipolazione strategica nelle raccomandazioni di notizie per ogni utente può intensificare i pregiudizi innati, portando alla segregazione ideologica e alla polarizzazione della società. Sebbene numerosi studi abbiano esplorato camere d'eco su varie piattaforme online, come blog, forum e social media, non è ancora stata affrontata una ricerca riguardante gli aggregatori di notizie. I lavori esistenti sulle piattaforme di social media, che enfatizzano fattori come l'omofilia e le relazioni sociali, non sono direttamente applicabili agli aggregatori di notizie a causa dell'assenza di tali relazioni tra gli utenti. Questo studio propone nuove misure quantitative e qualitative per caratterizzare le camere d'eco negli aggregatori di notizie attraverso diversi profili utente. Il nostro obiettivo è quello di offrire un contributo significativo, fornendo un quadro esaustivo della dinamica delle camere d'eco nel contesto del consumo mediatico. Tale approfondimento consentirà una comprensione più approfondita delle implicazioni sulla polarizzazione della società, esaminando e verificando l'esistenza degli ambienti virtuali attraverso esperimenti empirici.

Chapter 1

Introduction

The advent of social media and availability of news media platforms online has drastically transformed the news consumption behavior. Nowadays, the majority of the population tends to receive the daily news feeds online rather than the printed offline newspapers [Aral and Zhao, 2019]. Therefore, users have unbridled access to continuous stream of news information throughout the day and their attention is not bounded by the contents of any specific printed newspaper at hand. The transformation is marked by a 24/7 news cycle, diverse news sources beyond traditional outlets, customization and personalization through algorithms, increased interactivity and engagement on social media, incorporation of multimedia content for a richer experience, global reach facilitating access to international news, challenges posed by information overload, shifts in revenue models for news organizations, and the consequent impact on traditional print media with some outlets transitioning to digital formats. Additionally, the digital era has given rise to citizen journalism, empowering individuals to contribute to news reporting. This evolution underscores the dynamic nature of the media landscape, influencing journalism practices, audience engagement, and the financial sustainability of news organizations.

These impertinent changes in news consumption demand news media sources to understand the pulse of the readers and adapt according to the trends [Vrijenhoek et al., 2021]. Several studies indicate that users choose news media sources in a way that fulfill their specific communication needs. Therefore, to cater to the user needs, news media sources often align the news topics specifically to user interests and preferences [Diaz-Diaz et al., 2022, Levy and Razin, 2019]. This strategic approach aims to enhance user engagement and retention by delivering more relevant and appealing content, fostering an enhanced user experience. For example, a news aggregator app or site might prioritize articles on technology, international affairs, and en-

vironmental issues for a user with an history of reading such content. While personalization offers benefits, it also raises challenges, such as the risk of creating information bubbles or *echo chambers*. Existing research works [Ge et al., 2020, Jiang et al., 2019] define an *echo chamber* as an isolated virtual room or environment where the users' beliefs and interests are reinforced by repeated exposure to information that aligns with user's specific interests. Therefore, strategic manipulation and selective exposure in news recommendation coupled with user's confirmation bias can amplify the inherent biases, lead to ideological segregation [Flaxman et al., 2016] and polarization [Garrett, 2009] in society. This particular phenomenon is characterized by the selective exposure and reinforcement of individuals' pre-existing beliefs, opinions, and perspectives through the consumption of content that aligns with their worldview. This insular digital environment is perpetuated by algorithms, user's behaviors, and network structures that prioritize and amplify content in accordance with the user's past preferences and interactions. The result is a self-reinforcing feedback loop where users are continually exposed to content that reinforces their existing biases, creating a virtual space that shields them from diverse opinions and alternative viewpoints. Consider an individual with a specific political ideology who actively engages with and shares content from like-minded social media accounts. The platform's algorithm, designed to maximize user engagement, takes note of these preferences and subsequently tailors the individual's feed to predominantly showcase content that aligns with their political views. Over time, the user's social media experience becomes increasingly saturated with content that confirms their existing beliefs, as posts challenging or presenting alternative perspectives are deprioritized or filtered out. This individual, residing within the confines of the echo chamber, may find their opinions reinforced, while exposure to dissenting viewpoints diminishes, contributing to a heightened sense of polarization and a potential distortion of reality. The echo chamber effect in social media thus underscores the challenges of fostering diverse discourse and critical thinking within digital spaces. Linked to filter bubbles, these personalized information ecosystems may result in political and ideological polarization, reduced empathy, and the erosion of objective truth. This phenomenon has implications for decision-making processes, as individuals may be influenced by skewed information, and it poses challenges in combating manipulation and misinformation.

There are several research works which focus on understanding and identifying echo chambers in different platforms, such as, blogs [Gilbert et al., 2009], forums [Edwards, 2013] and social media platforms [Zannettou et al., 2018]. These research works focus on different objectives with respect to echo chambers, such as, understanding of the factors that lead to formation, charac-

terization and identification of echo chambers, its impact on polarization in society [Baumann et al., 2020, Cinelli et al., 2020] or provide automated solutions to prevent formation of echo chambers. For example, Cinelli et al. [Cinelli et al., 2020] and Wang et al. [Wang et al., 2020] study the role of homophily in user interactions, cognitive biases and similar political ideology as major factors that leads to echo chambers. Additionally, several research works highlight the significance of social network characteristics, such as, signed networks [Garimella et al., 2018], follower followee relationships [Duseja and Jhamtani, 2019], temporal relationships [Morini et al., 2021], random walks, balance theory [Interian et al., 2023] and community structure [Cossard et al., 2020, Morini et al., 2021], to identify echo chambers. Ge et al. [Ge et al., 2020] explore formation in echo chambers specifically in e-commerce recommender systems on the basis of variance in user's recommended products on the basis of content diversity. Additionally, Sasahara et al. [Sasahara et al., 2021] propose several mechanisms to mitigate echo chambers. However, these existing recent research works focus on identifying echo chambers on different social media platforms and there is no research work that focuses on echo chamber in news media aggregators. Existing works on social media platforms are not directly applicable for news media aggregators as there is no social relationship among users and users have only information about the news articles recommended to them. Unlike social networks where interactions and relationships are central, news aggregators focus primarily on delivering news content, with users seeking information rather than social engagement. News aggregators lack the extensive user profile information used in social media, and their recommendation algorithms are centered around news preferences rather than social connections. User engagement on news aggregators is measured through interactions with news articles, reflecting a distinct set of behaviors from social media platforms. While both platforms may utilize recommendation algorithms, the fundamental design and user experience considerations differ due to the varying purposes and intentions of users in consuming news versus socializing. Privacy considerations also diverge, with news aggregators generally dealing with simpler privacy models compared to the intricate privacy concerns of social media. Therefore, existing metrics based on homophily or user relationships which have been proven to be highly effective in identification and studying of echo chambers in social media platforms are inapplicable to news media aggregators. In this study we propose different measures that can quantitatively and qualitatively study characterization of echo chambers in news media aggregators across different users.

Chapter 2

News Media Ecosystem and Context

In the digital age, the landscape of news dissemination has been further transformed by the emergence of news media aggregators [Angela M. Lee, 2015]. These platforms have brought many websites and apps where users can quickly find news articles selected just for them, based on what they like.

However, it's not just users who reap the benefits of these aggregators. Websites that permit the publication of their articles also stand to gain significantly. By being featured on news media aggregators, these sites can broaden their reach and visibility, attracting a larger audience and driving traffic to their platforms. As news media aggregators continue to grow in prominence, so too do the opportunities for websites to expand their readership and enhance their online presence.

2.1 News Media Aggregators

A *news media aggregator* is a digital platform or service that collects and compiles news articles, updates, and content from various sources across the internet into a single location. These aggregators use algorithms to organize and present the information to users based on their preferences and interests [Aiello et al., 2012, Kossinets and Watts, 2009]. Rather than creating original content, news media aggregators aggregate articles and news pieces from different publishers, providing users with a convenient way to access a wide range of news sources and topics in one place. News media aggregators provide several advantages for both users and publishers: users benefit from the convenience of accessing news from multiple sources in one place, saving time and effort. These platforms offer a diverse range of content, allowing

users to explore various perspectives and topics [Susan Athey, 2021]. Personalization features further enhance the user experience by tailoring news feeds to individual interests. For publishers, being featured on a news aggregator can significantly increase visibility and drive traffic to their websites, potentially attracting new readers. Some aggregators also offer monetization opportunities, allowing publishers to generate revenue through advertising or subscription models.

In this study, we underscore the significance of providing users with a personalized space that automatically populates with news articles relevant to their interests [Del Vicario et al., 2016]. This approach ensures that users are presented with content that truly matters to them, enhancing their experience and engagement with the platform. Moreover, in instances where the curated articles may not fully meet the user's preferences or needs, the homepage of the news aggregator serves as a dynamic gateway to explore new and diverse news topics. This feature ensures that users have access to a continuous stream of fresh content. By prioritizing user-centric design, news aggregators can effectively cater to the diverse needs and preferences of their audience, maximizing user satisfaction and retention. Different online sites and newspapers often present news using distinct writing styles and language choices. These nuances in language usage have a significant impact on the interpretation and perception of the information, particularly within specific contexts and topics. The primary focus of this study is to investigate how news aggregators curate and deliver news content to users based on their search history, reading habits, and any predefined preferences set by the aggregator. This phenomenon is commonly referred to as bias within news media aggregators. Bias in news aggregation occurs when the aggregator selectively presents or prioritizes certain news articles over others, potentially influencing the user's perception of the news landscape. The goal of this study is to ascertain whether such biases exist within news aggregators and, if so, to identify the direction of these biases. By examining the content selection and presentation practices of news aggregators, we aim to uncover any patterns of imbalance or favoritism towards specific viewpoints, ideologies, or sources.

2.1.1 Google News

Choosing Google News as the focus of our study offers several compelling reasons. It boasts a user-friendly interface, seamlessly integrating multimedia elements for a comprehensive news consumption experience. Its global reach ensures exposure to international perspectives, while the efficiency in news aggregation categorizes content for easy exploration. Google News prioritizes

trustworthy sources, contributing to its credibility, and for users within the Google ecosystem, the platform offers seamless integration.

We also considered other news media aggregators, including Yahoo Answers, Apple News, Smartnews, News360 and Flipboard, but we opted for Google news because it provides well-documented support for APIs. The following features are essential for our analysis:

- **Homepage**: the Google News homepage serves as a global snapshot, featuring the most popular and significant news stories from around the world. It display the most popular and significant news stories from across the globe, and it provides a quick overview of global events, covering a diverse range of topics to keep the user informed about what's happening internationally.

- **For You Page**: the For You Page is a personalized section for individual user interests. Drawing insights from users' past news interactions, it curates a selection of articles focused on specific topics that align with their preferences. This feature aims to enhance user engagement by delivering content that caters to their unique interests.

- **Section of Various Topics (customizable)**: Google News offers a range of pre-defined sections covering diverse topics such as World, Local, Business, Technology, Entertainment, and Sports. Users can customize these sections based on their preferences, ensuring a personalized news feed that aligns with their specific interests.

Our approach involve engaging the users into the Google News ecosystem, where, upon logging in, they encounter a personalized area of news articles day by day. This deliberate curation result in the formation of distinct personal networks of news—a collection of information finely tuned to the preferences and assigned topics of each user.

2.1.2 News Articles

An integral component of the subsequent analyses, as well as a fundamental parameter of this study, revolves around how the news articles are presented by the news media aggregator. These articles serve as the primary source upon which our examination is based. We closely examine different aspects of the articles, looking at both how they're written and what they're about. Firstly, we analyze the topical focus of each news article, which may encompass a spectrum of subjects ranging from political topics and global events to more leisurely topics such as sports or entertainment. Secondly, we explore

how these subjects are explained in the articles. This entails a comprehensive assessment of the linguistic choices employed by the authors, including vocabulary selection, tone, and stylistic elements. Moreover, we explore the syntactic structure of the sentences that constitute the news articles, evaluating how words are employed within the context to convey meaning and perspective. By meticulously dissecting these key elements of the news articles, our aim is to unearth any discernible patterns or biases in the way information is presented and framed by the news media aggregator.

Suppose we analyze two different sentences, which they're using the same words, but written in different ways. The sentence could be: "Non-stop progress and prosperity are the hallmarks of the visionary policies enacted by the X[1] party, ushering in an era of unparalleled growth and opportunity for all citizens". This sentence is presenting a positive viewpoint towards the X party, suggesting that their policies have led to continuous progress and prosperity. It portrays the party's actions as visionary and beneficial for the growth and well-being of the nation. At the other hands, we could find: "Ah, yes, because clearly, non-stop progress and prosperity are just magically raining down on us from the sky, thanks to the brilliant policies of the X party". In this sentence, the tone shifts to one of irony and sarcasm. The same words and phrases are used as in the first sentence, but the intention is to mock or criticize the idea presented. The use of sarcasm implies skepticism or disbelief towards the notion that progress and prosperity are solely attributed to the policies of the X party.

2.2 Political Bias in News Media

Political topic news are presented every day by any newspaper and any aggregation site: they inform citizens about the actions and decisions of their governments, helping them make informed choices during elections and hold their leaders accountable for their actions [Pye, 2015]. They also provides insights into international relations, diplomacy, and geopolitical events, helping people understand the interconnectedness of the world and the implications of global events on their own countries. It is inevitable that greater importance is assigned to this type of topic. Political bias in news articles can contribute to the formation of echo chambers [Nathan Honeycutt, 2023]: when news consistently present information from a particular political perspective, they reinforce the beliefs of their audience while potentially ignoring

[1]In this context, "X party" is a generic placeholder term and does not refer to any real or specific political party. It is used for illustrative purposes only.

or downplaying alternative viewpoints. Political news can more easily create echo chambers compared to other topics due to several factors: political beliefs often form a core part of individuals' identities, and people tend to be more emotionally invested in political issues, which can lead them to seek out news sources that confirm their existing beliefs, reinforcing their preconceived notions. Also, political discourse is frequently polarized, with issues framed in terms of "us versus them" or "right versus wrong." This polarization can make it challenging for individuals to consider alternative viewpoints, leading them to gravitate towards news sources that align with their own ideological leanings. Social media platforms, as we said, often use algorithms to personalize users' news feeds based on their past interactions and preferences. This can logically create Filter Bubbles, where individuals are exposed only to content that reinforces their existing beliefs, further entrenching echo chambers.

To ensure a comprehensive examination of the creation of echo chambers across diverse cultural landscapes, we expand our study to encompass two geographically distinct countries: USA[2] and India, respectively. By selecting these two nations, each with its unique political ideologies, cultural contexts, and societal structures, we aim to capture a broad spectrum of perspectives and factors contributing to the formation of echo chambers. This choice allows for a comparative analysis of political discourse and news consumption patterns in countries with contrasting historical backgrounds, political systems, and media landscapes. For USA, we direct our attention to two distinct branches of political news: the Republican and the Democratic Party [Freeman, 1986]. By examining the coverage of these two major political parties, we aim to gain insights into the differences in news presentation by Google News. The Republican Party, often associated with conservative ideologies, tends to prioritize topics such as limited government intervention, fiscal responsibility, and traditional social values in its news coverage [Gienapp, 1856]. Conversely, the Democratic Party, known for its progressive ideologies, tends to prioritize topics such as social justice, environmental protection, healthcare reform, and economic equality in its news coverage [Saribay, 1960]. Similarly, in the context of India, news coverage often exhibits distinct biases depending on whether it aligns with the ruling government or the opposition parties. Pro-Government news tend to highlight the achievements, policies, and initiatives of the ruling party, portraying them in a positive light. Conversely, Pro-Opposition news may adopt a more critical stance towards the government, highlighting its shortcomings, failures, and controversies. These outlets may use language that critiques gov-

[2]United States of America.

ernment policies, highlights instances of corruption or mismanagement, and advocates for alternative approaches proposed by opposition parties. Given the numerousness and difference of the topics covered by both parties, the language used in news articles can significantly influence the way information is perceived and interpreted, potentially contributing to the creation of echo chambers. The choice of words, tone, and framing employed by news outlets can subtly shape readers' perceptions of political events and individuals associated with each party. The use of such language can subtly reinforce existing beliefs and biases, leading readers to gravitate towards news sources that align with their political views while disregarding or dismissing alternative perspectives. Comparing the news content associated these parties allow us to discern patterns of topic selection, emphasis, and framing within each political sphere.

In this thesis, the primary focus is on the perspective from USA, as the analysis primarily revolves around news articles from this region. The decision to concentrate on the USA stems from my direct involvement in gathering and analyzing data from american news sources. However, while the thesis is focused on the USA context, insights and findings from India are also presented and discussed.

2.2.1 Historical Context and Influences

The chronological snapshot of the start of 2023 in politics reveals a period marked by significant events that undoubtedly influenced the dynamics of our study. Clearly, every analysis that involves experiments and the collection of data from temporal sources inherently carries a specific temporal timestamp. This timestamp denotes the period during which the project is conducted and the data is gathered. Importantly, this temporal aspect is not merely a logistical consideration; rather, it profoundly influences the nature and interpretation of the data collected. To illustrate, let us consider the example of extracting data from the USA political landscape during a significant event such as an impeachment proceeding against a sitting president, as occurred with President Trump in August 2023. In such a scenario, the data captured reflects the prevailing sentiments and dynamics surrounding this event. For instance, individuals' attitudes, behaviors, and interactions on social media platforms or news websites may be influenced by the unfolding events related to the impeachment. As a result, the data collected during this period may inherently carry a bias, reflecting the heightened emotions, polarization, and discourse surrounding the impeachment proceedings. Users' engagement with political content during this time may be skewed towards viewpoints either in support of or against the president's indictment,

depending on various factors such as their political affiliations, ideological leanings, and media consumption habits.

This temporal bias underscores the importance of contextualizing and interpreting data within the broader socio-political landscape of the specific time period in which it was collected. Failure to account for these temporal dynamics can lead to erroneous conclusions and misinterpretations of the data. Therefore, researchers and analysts must be cognizant of the temporal context surrounding their data collection efforts and exercise caution when drawing conclusions or making inferences based on temporally bound datasets. By acknowledging and accounting for these temporal influences, researchers can ensure the accuracy, validity, and reliability of their findings in projects involving temporal data sources. Notably, the context was shaped by the indictment of Donald J. Trump, a development that coincided with a vote to authorize an impeachment inquiry into him. Against this backdrop, the political landscape experienced a seismic shift with the expulsion of Rep. George Santos (R-N.Y.)[3] from Congress due to an Ethics Committee investigation. Furthermore, the power dynamics within the House GOP underwent a historic transformation, marked by the ousting of Speaker McCarthy and the subsequent election of Mike Johnson (R-La.)[4]. The geopolitical arena also played a role. Simultaneously, domestic elections, including those in Kentucky, Virginia, and Ohio, saw significant victories for Democrats and abortion rights activists. In a surprising turn of events, Sen. Joe Manchin (D-W.Va.)[5] announced he wouldn't seek re-election and hinted at a potential third-party presidential run.

These real-world occurrences undoubtedly influenced the subjects and themes under examination in our study, adding layers of complexity to the analysis of news consumption behaviors during this historically significant period. By tailoring our study to the news consumption behaviors, I intentionally center our examination on the intricacies of the USA political landscape and media dynamics. While my direct involvement pertained to managing USA, the study of news about India adds a valuable comparative dimension, broadening the scope of our analysis. The historical events and contexts within India during the specified period are integral to understanding the dynamics of news consumption among users from India. These factors, whether political, social, or economic, contribute to the intricate tapestry of media interactions, influencing the topics covered, the tone of reporting, and the overall news landscape.

[3]Republican - New York.
[4]Republican - Louisiana.
[5]Democratic - West Virginia.

Chapter 3

Related Work

Understanding the phenomenon of echo chambers within the news media aggregators necessitates a comprehensive exploration across diverse digital platforms [Thompson and Santos, 2023]. This section delves into existing research and analyses, shedding light on the multifaceted nature of echo chambers and their manifestations in various online environments. Through an examination of echo chambers in forums, blogs and other digital spheres, this section aims to contextualize the phenomenon within broader scholarly discussions and elucidate its implications for information dissemination, societal discourse, and individual perspectives.

Online journals have emerged as significant arenas for the exchange of ideas, opinions, and information. Within these virtual spaces, users congregate around shared interests, forming communities that often reinforce existing beliefs and perspectives [Aiello et al., 2012, Kossinets and Watts, 2009]. Different studies have explored the dynamics of echo chambers within online forums, examining how group dynamics, moderation practices, and algorithmic recommendations contribute to the proliferation of homogeneous viewpoints. By elucidating the mechanisms underlying echo chambers in online sites, researchers have sought to discern the consequences for information diversity, ideological polarization, and the formation of collective identities.

3.1 Echo Chambers in Social Media

The debate surrounding social media echo chambers is multifaceted and complex. On one hand, proponents argue that digital technologies and social media platforms have democratized access to information, facilitating the formation of diverse online communities and empowering individuals to engage in public discourse. They contend that these platforms provide a platform for

marginalized voices to be heard and foster a more inclusive and participatory public sphere. Conversely, critics warn that social media echo chambers exacerbate ideological polarization and contribute to the fragmentation of public discourse [Ludovic Terren, 2021]. By algorithmically curating users' content feeds based on their past behavior and preferences, social media platforms inadvertently reinforce existing biases and filter out dissenting viewpoints [Cinelli et al., 2020]. This selective exposure to information can create echo chambers wherein individuals are exposed primarily to content that reaffirms their preconceived beliefs, further entrenching ideological divides and hindering constructive dialogue.

The scientific literature on social media echo chambers reflects this complexity, with studies employing a range of methodologies and approaches to investigate the phenomenon. Some studies have utilized digital trace data to analyze patterns of information consumption and sharing on social media platforms, uncovering evidence of echo chamber effects. Others have relied on self-reported data to gauge individuals' perceptions of echo chambers and their impact on their online behavior. Moving forward, it is essential for researchers to recognize the limitations and biases inherent in different methodological approaches and to explore innovative ways of combining self-reported and digital trace data to gain a more comprehensive understanding of social media echo chambers [Daejin Choi, 2020]. By doing so, scholars can contribute to a more nuanced and empirically grounded understanding of the impact of social media on democratic processes and public discourse, informing policy interventions and platform design aimed at mitigating the adverse effects of echo chambers while preserving the benefits of digital connectivity and information sharing.

3.2 Echo Chambers in Blogs

An online blog, short for "weblog," is a digital platform where individuals, groups, or organizations can publish and share content in a structured and chronological format [Kim, 2008]. Rooted in the democratizing ethos of the internet, blogs have evolved into versatile tools for self-expression, information dissemination, and community building. Blogs often feature a distinctive authorial voice, reflecting the unique perspectives, experiences, and expertise of the individual or group behind the content. This personal touch distinguishes blogs from traditional media outlets, imbuing them with authenticity, relatability, and a sense of intimacy that resonates with audiences.

In the dynamic landscape of online discourse, blogs stand as prominent platforms where individuals engage in the dissemination and consumption

of information, opinions, and narratives. Within this virtual realm, the phenomenon of echo chambers manifests itself as individuals gravitate towards blogs that align with their preexisting beliefs, values, and ideological inclinations [E. Gilbert and Karahalios, 2009]. Echo chambers in blogs are characterized by the formation of insular communities wherein like-minded individuals congregate, reinforcing and amplifying shared perspectives while marginalizing dissenting viewpoints. At the heart of echo chambers in blogs lies the intricate web of interpersonal relationships and social dynamics that underpin online communities. Blogs serve as virtual spaces where individuals from diverse backgrounds converge to share their thoughts, experiences, and perspectives on different discussion topics ranging from politics and culture to personal hobbies. As users interact with blog content through comments, likes, and shares, they forge connections with like-minded individuals who resonate with their beliefs, values, and interests. These interactions foster a sense of camaraderie and belonging within the blogosphere, cultivating a community of individuals united by shared affinities and mutual understanding. However, within the seemingly inclusive confines of these online communities, hides the insidious phenomenon of echo chambers. As individuals immerse themselves in the echo chamber of their chosen blogosphere [Wolfowicz et al., 2023], they unwittingly insulate themselves from dissenting viewpoints and alternative perspectives. This process of ideological reinforcement is fueled by a collective tendency to prioritize affirmation over interrogation, consensus over dissent, and conformity over critical engagement.

To address the core issue of echo chambers in blogs, users need to raise an environment where everyone values open-mindedness, thoughtful discussion, and mutual respect online. This involves welcoming diverse viewpoints, engaging in constructive conversations, and being open to exploring new ideas.

3.3 Echo Chambers in E-commerce

In the realm of e-commerce, the advent of personalized recommendation systems has revolutionized how users navigate and interact with online platforms [Tassabehji, 2003]. These systems leverage advanced algorithms to tailor product recommendations to individual user preferences, enhancing the efficiency and effectiveness of content discovery. However, amidst the benefits of personalized recommendations lies the growing concern about echo chambers within e-commerce platforms.

It refers to the phenomenon wherein users' interests are reinforced through repeated exposure to similar content. This self-reinforcing cycle is perpet-

uated by personalized recommendation algorithms, which prioritize items that align with users' past preferences and browsing history. As users engage with these recommendations, they are increasingly directed towards a narrow subset of products that mirror their existing tastes and preferences. The implications of echo chambers in e-commerce are twofold. On one hand, they can enhance user satisfaction and streamline the shopping experience by presenting users with relevant and appealing products. By tailoring recommendations to individual preferences, e-commerce platforms can increase user engagement, conversion rates, and overall satisfaction [Yingqiang Ge, 2020]. However, the proliferation of echo chambers also raises concerns regarding information diversity, consumer choice, and market dynamics. By limiting users' exposure to a narrow range of products and brands, echo chambers may inhibit serendipitous discovery, impede market competition, and reinforce existing consumer biases. Moreover, they can exacerbate Filter Bubbles, wherein users are insulated from alternative perspectives and product offerings, further entrenching their existing preferences [Alatawi et al., 2023].

To address these challenges, researchers and practitioners are increasingly focusing on understanding and mitigating the impact of echo chambers in e-commerce by analyzing user behavior data, evaluating recommendation algorithms, and implementing interventions to diversify content exposure. Insights gleaned from such analyses can inform the refinement of recommendation algorithms, the design of user interfaces, and the development of platform policies aimed at promoting information diversity. Through concerted efforts, stakeholders can mitigate the adverse effects of echo chambers while harnessing the benefits of personalized recommendations to enhance the online shopping experience for users worldwide.

Chapter 4

Data Gathering and Experiment Design

To study real life users news consumption behavior without any inherent bias, we carefully design our simulation such that we have users with different news topical interests who belong to different locations. The reason being users can have varied news topical interests and the formation of echo chambers can depend on the basis of the users topical interests and even locations. Therefore, we consider different news topic interests, selected on the basis of their popularity in Google News platform.

4.1 User Profile Creation

For our experiments, we conduct user profile creation task from both USA and India. To ensure a comprehensive representation of user demographics and preferences, we divide the user creation process among the team members working on the project, in which I actively engage representing a portion of USA users. This approach allows us to capture diverse perspectives and insights from individuals hailing from distinct cultural backgrounds and geographical regions.

We consider 38 users such that 18 users belong to USA and 20 users belong to India. Originally, we had two additional users belong to USA, as a total of 40 users. However, as part of our attempt to simulate their physical presence in the same location for experimental purposes, we ran into an unexpected obstacle. To achieve this, we utilized a VPN[1] to ensure both users appeare to be accessing Google News from the same geographical location. Regrettably,

[1]Virtual Private Network: an arrangement whereby a secure, apparently private network is achieved using encryption over a public network.

the VPN, designed to enhance privacy and security, inadvertently triggered Google News to detect unusual changes in the users' locations. The dynamic nature of the VPN, which altered the region within the USA periodically, leds to the suspension of our accounts on Google News due to potential security concerns. Despite these challenges, we proceeded with the study, focusing on the remaining users within the constraints imposed by Google News policies. We avoided using the VPN, and limiting ourselves to setting the localization of the news within Google News to USA (for users from USA) and India (for users from India). Opting to study news consumption in two different countries adds valuable dimensions to our research. The cultural diversity inherent in examining two distinct contexts provides insights into how societal values shape echo chambers. Variability in political landscapes, language dynamics, and media environments between the chosen countries allow for a comprehensive exploration of how these factors influence the prevalence and characteristics of echo chambers. Additionally, the study's global perspective fosters cross-cultural learning, enabling a nuanced understanding of news consumption behaviors.

4.1.1 User Characteristics

We adopt a systematic approach to user identification by assigning a unique ID to each user. Specifically, we categorize users based on their geographical location, distinguishing between those from the USA (U) and those from India (I). For USA users, we designate IDs ranging from U_1 to U_8 (handled by me), and from U_9 to U_{18} . Similarly, users from India are assigned IDs from I_1 to I_{20} . To ensure a thorough examination of our research, we include participants from diverse backgrounds. Our approach involves creating individuals aged 20 to 70 years, representing a wide range of age groups. Additionally, we aim for gender balance by including an equal number of male and female participants. Moreover, to capture varied perspectives, we enlist participants born in both European and non-European countries.

By incorporating participants with such diverse characteristics, we're better equipped to investigate whether Google News treats users differently based on different human factors, and we can gain insights into potential biases or disparities in the platform's content delivery and user experience. However, extracting information regarding the possibility that the platform creates bias regarding a single user characteristic is difficult, but nevertheless it gives a general idea of the fact that it can actually create differences regarding different aspects of user creation. For generating name and surname of users, we turn to websites that create made-up name combinations. These sites provide fictional first and last names that aren't

linked to real people. This approach ensures that the names we use for our users are entirely fictitious and don't belong to any actual individuals. We assign an email to users to access Google News created by combining *<name><surname>@gmail.com*.

4.1.2 Topic Assignment

We assign each user two main news topics. The first topic, called *majority topic*, is the one that the user is most interested in. This means that they read more news articles about this topic compared to the second assigned topic. On the other hand, the second assigned topic is referred to as the *minority topic*. Even though users read news about both topics every day, they read fewer articles about the minority topic compared to the majority one. This setup allows us to examine how users engage with news content when one topic is prioritized over another, helping us identify any biases or disparities that may arise in Google News recommendations. By adopting this approach, we can analyse how Google News ranks topics within its platform. Additionally, we establish a chronological sequence of news articles read by each user. This allows us to investigate whether Google News consistently delivers additional news on specific topics in the days following, based on the user's past reading habits.

Existing research works on echo chambers in social media platforms have highlighted that the echo chambers characterization can vary on the basis of the user's political leaning. Therefore, we consider users with three different types of political leaning: Republican political leaning, Democratic political leaning and Neutral political leaning, for USA. Subsequently, for India, we consider the political leaning as Government leaning, Opposition leaning and Neutral political leaning. We choose to assign a specific political party interest to the first 6 users for each group: users U_1 to U_6 and U_{11} to U_{16} for USA, and users I_1 to I_6 and I_{11} to I_{16} for India. We prioritize users with an interest in politics because, as previously mentioned, political news often carries inherent biases, with conflicting viewpoints and the formation of like-minded groups. By focusing on these users, we aim to investigate whether Google News exhibits any biases towards or against a particular political party. Our goal is to understand whether Google News tends to provide news favorable to a specific political party to a specific user, or if it exhibits biases by presenting news in favor of the opposing party. We're particularly interested in observing whether Google News adjusts its news recommendations based on users' past reading history, potentially reinforcing existing biases or presenting news from alternative perspectives. Through this investigation, we hope to gain insights into the mechanisms underlying

news delivery on Google News and its implications for political discourse and information dissemination.

However, we consider different news topic interests, such as, Politics, Sports, Entertainment, Technology, World and Business. Although there can be other news topics, we select these topics on the basis of the popularity of the topics in Google News platform. A summary of the users with their news consumption behavior for both the the macro and micro topic, is provided in Table 4.1.

4.1.3 Chronological Politic Affiliation

When a user is assigned the Republican Party as their primary topic, it means they not only read news related to the Republican Party of the USA, but also news that favors the Republican Party. This ensures that the user is exposed solely to news content aligned with their political affiliation, thus creating a personalized history consisting exclusively of news favorable to the Republican Party. This approach allows us to closely examine Google News behavior through detailed analyses of the article content provided. We aim to determine if Google News consistently delivers articles pertaining only to a specific political party, and moreover, if these articles are biased in favor of that party. Similarly, users assigned to the Democratic Party in the USA, as well as users assigned to pro-Government and pro-Opposition parties in India, follow the same pattern. By focusing on users with specific political affiliations and analyzing the type of news content provided, we seek to understand how Google News behaves and whether it exhibits any biases towards particular political parties.

A user considered neutral is the user who reads news regarding any political party, within which there is no evident bias towards one political party, compared to another, or who belittles one, bringing the other to the center of the discussion. This path applies to both the USA and India.

Table 4.1: Table shows the News Consumption Interests of Users.

Location	Users	Macro	Micro	Location	Users	Macro	Micro
	U_1	Republican	Music		I_1	Government	Sport
	U_2	Republican	TV		I_2	Government	Entertainment
	U_3	Democratic	Entertainment		I_3	Opposition	Sport
	U_4	Neutral	Tennis		I_4	Opposition	Entertainment
	U_5	Science	Animals		I_5	Neutral	Others
	U_6	Football	Movies		I_6	Popular	Popular
	U_7	Movies	Fashion		I_7	Sport	Neutral
USA	U_8	AI	Food	$India$	I_8	Entertainment	Neutral
	U_9	Democratic	Sport		I_9	Technology	World
	U_{10}	Democratic	Entertainment		I_{10}	World	Business
	U_{11}	Republican	Sport		I_{11}	Government	Sport
	U_{12}	Republican	Entertainment		I_{12}	Government	Entertainment
	U_{13}	Neutral	Others		I_{13}	Opposition	Sport
	U_{14}	Popular	Popular		I_{14}	Opposition	Entertainment
	U_{15}	Sport	Pol		I_{15}	Neutral	Others
	U_{16}	Entertainment	Pol		I_{16}	Popular	Popular
	U_{17}	Technology	World		I_{17}	Sport	Neutral
	U_{18}	World	Business		I_{18}	Entertainment	Neutral
					I_{19}	Technology	World
					I_{20}	World	Business

4.2 Retrieving News

In crafting our dataset, we meticulously simulate the user experience of retrieving news on Google News. As an initial step, given the absence of pre-existing news in the For You Page, we initiate the process by actively searching for specific topics using the search bar. This approach allow us to mimic the real-world scenario where users actively seek out news content aligned with their interests.

4.2.1 Simulating Human Behavior

Once the relevant news articles are retrieved, we simulate user behavior by engaging and interacting with the content of the article. This encompassed actions such as clicking on news articles, initiating the reading process, and simulating the natural behavior of scrolling up and down to explore the entire article. To further mimic authentic user engagement, we implement random highlighting of text, emulating the common practice of user reading an article. Crucially, we incorporate the element of time into our simulation. Users are simulated to stay on a news page for an approximate duration of five minutes. This deliberate duration aim to emulate a user genuinely investing time in comprehensively consuming the news content, signaling to the algorithm that the topic was of substantial interest. This methodology aim to ensure that the algorithm could accurately discern user interest and prefer-

ences, thereby contributing to a more authentic and informative analysis of news consumption behaviors.

Over the course of several days, each user click on a total of five news articles daily, with a distribution of three articles focusing on macro topics and two on micro topics. This intentional repetition of user interactions is designed to enable the algorithm to progressively learn and discern the individual interests of each user. As users consistently engage with news content across both macro and micro topics, the algorithm adapt and begin to refine its understanding of their preferences. Through this iterative process, a notable evolution occurs in the algorithm's functionality. Over time, the For You Page, initially devoid of pre-existing content, transform into a personalized space. The algorithm, leveraging insights gained from users' repeated interactions, starts populating the personalized user area with news articles aligned with his specific interests. This encompassed both macro and micro topics, reflecting a more nuanced and tailored representation of the users' news preferences.

4.2.2 Dataset

In essence, our approach seek to illustrate how consistent user engagement on Google News over an extended period facilitate the algorithm's learning process. The outcome is a dynamically evolving For You Page that increasingly mirror the diverse interests of each user, demonstrating the platform's capacity to adapt and enhance the user experience based on individualized content preferences. Table 4.2 presents a snapshot of the head of our dataset, providing the essential information collected during our simulated user interactions on Google News. The dataset encapsulates details about news articles, necessary for our analysis, including the *user ID, topics, news ID, title, description* of the news, the corresponding *link*, and the *date* of news retrieval. Originally, we also sought to include information about the physical location of the simulated user, reflecting the geographic location of the device used during the news collection process, and the time at which the news was retrieved. However, upon meticulous examination, we observed that these two parameters did not contribute substantially to our research objectives. Our dataset encompasses approximately 100 news articles about macro topic, and half about micro, collected over a period of two months, specifically in March and April. It is important to note that the dataset reflects an extended simulation process that started around October and November of the preceding year. The simulated user interactions, involving the reading and clicking of news articles, were initiated during the mentioned period. When we officially started collecting news articles information to our dataset, Google News had

already been learning for a while. The algorithm had started to understand and adjust to the pretend interests of the users, getting better at suggesting articles based on how users were interacting with them. This temporal aspect is crucial to understanding the contextual backdrop of our dataset. The two-month period in which we officially collected news articles serves as a consolidated snapshot of the ongoing evolution in user preferences and algorithmic adjustments that had been taking place since the initiation of the simulation in the preceding months.

4.2.3 User Engagement with Pro-Topic Articles

During our simulation, when we come across news articles that didn't seem to match the interests of our simulated user (especially in the complicated area of politics), or when the title of the article is simply ambiguous, we take a careful approach. If a headline seems unclear or misleading, we don't immediately read the article using the simulated user session. Instead, we first check to see if the content matches what our user would be interested in. Many articles, especially those dealing with political topics, are difficult to interpret without reading them first. Consequently, we don't know if it's pro or against a particular topic (which we would have liked to add to the history of a particular user who was reading an article about a particular topic). To achieve this, we open an additional browser without an active user session. This parallel browsing approach allows us to access and preview the news article from a neutral standpoint before officially registering it as "retrieved" within the logged-in session. By doing so, we can discern the nature and tone of the news article; this practice provides a valuable feedback loop. It enables us to make informed decisions about whether to proceed with the article in the simulated user session, ensuring a more accurate representation of the user's engagement with content that genuinely resonated with their interests.

In addition to determining which articles to present to the user, we implement techniques to ascertain which of these articles were worthy of actual reading. This consideration becomes particularly crucial for political topics, where the goal is not only to address the subject matter but also to discern if the article exhibited a favorable stance toward the topic. To achieve this, we deploy various Natural Language Processing (NLP) techniques, including Sentiment Analysis, Readability Analysis, and others. These techniques allow us to calculate different scores for each article, forming a multidimensional assessment beyond the topic coverage. The metrics considered aspects such as sentiment, the complexity of language, and other readability factors. Crucially, we establish thresholds for these scores, defining a level of

"goodness" that an article needed to surpass. If the goodness score exceed our predetermined threshold, it indicate not only that the article addressed the topic but also that it presented a favorable perspective. The detailed exploration of these analyses will be presented in the subsequent chapter, providing a comprehensive understanding of the criteria employed to determine the quality and stance of the articles within the context of our research. Figure 4.1 delineates the interaction reading flow within our reading framework, encapsulating the dynamic engagement between users and articles.

4.2.4 Automated Collecting Data System

Collecting news for our study prove to be a time-intensive process, particularly when simulating the daily routines of multiple users engaging with diverse news articles. This task becomes even more pronounced when we need to repeat the processfor approximately 10 users for each component of the project, each requiring distinct news content for their daily interactions. One significant time-consuming aspect is the meticulous attention to detail in simulating the entire user experience. This involves not only the actual reading of news articles but also the additional actions that real-world behavior on interactive elements. These actions are crucial for creating a realistic dataset that accurately reflected the diversity of user engagement with news content. Furthermore, factoring in the time requires for each simulated user to log in, navigate through the platform, and engage with the news articles introduced additional layers of complexity. Cumulatively, these tasks make the daily time investment for the simulation process quite substantial.

Building a unique dataset on which to carry out research, especially if personal, it takes time, especially because it's built from scratch. At first, we planned to study echo chambers in online news aggregators on a larger scale worldwide. However, we faced a challenge in gathering enough material for a thorough study. As we looked into collecting data, we realized that Google News alone offered a lot of material for our research. To prepare for future versions of our research and analyses, we set ourselves up strategically to avoid manually reconstructing the dataset. We started creating an automated system for collecting news. This system streamlines the process of gathering data, saving us from the tedious job of compiling it manually. In the beginning, we start automating our process by studying the API documentation provided by big news aggregators like Apple and Yahoo. Yahoo's documentation is especially helpful, as it give us clear instructions on how to integrate their system into a web app. Leveraging Python's extensive library ecosystem, we identify Selenium as a powerful open-source framework for automating web browsers. Selenium, compatible with various

web browsers such as Chrome, Firefox, and Safari, empowers developers to write scripts, including those in Python, to automate interactions with web browsers. Our approach involves the implementation of a straightforward system within Python, where Selenium is utilized to programmatically control web browsers. This system facilitate the simulation of user interactions with news articles, automating tasks such as clicking buttons, navigating pages, and extracting data. Our system exploit Selenium to simulate user login, providing a username and password corresponding to one of the simulated users. This allow us to seamlessly automate the process of accessing news articles significantly streamlining the data collection process for future iterations of our research. Once logged in, our automated system seamlessly navigates to the Yahoo homepage, where news content is presented to the user. While the content itself differs from that of Google News, the underlying approach remains identical. Leveraging Selenium's capabilities, which enable the identification of elements on a web page through various methods, we replicate the user's manual interactions in an automated fashion. Selenium's capacity to locate elements on the web page, such as those containing specific strings matching the topics we provided, facilitate the simulation of the user's manual click. This automated interaction ensure that our system could efficiently access news articles aligned with the predefined parameters, emulating the user's engagement with the content.

The main part of our automated system is figuring out what kind of news a user really likes. For example, if the system is set up for someone who supports the Republican Party, it's tricky to tell which news articles match their actual interests and which ones don't. By assigning specific attributes to the system, we enable the identification of objects within the web page, such as news links, that correspond to a designated string, in this case, "republican." This results in a collection of news links encompassing content both in favor of and against the Republican Party. However, the system, operating solely on the provided string, lacks the autonomy to independently distinguish between articles supporting and opposing the user's preferences. This is where the integration of Natural Language Processing analyses becomes essential. The NLP analyses, including Coherent Topics, Sentiment Analysis, and Latent Dirichlet Allocation (LDA), addressed in the next chapter, enables us to assign a goodness score to the text of a news article. This score acts as an indicator of the article's alignment with a particular topic. In practical terms, for a pro-Republican user, we run these parallel analyses during navigation. The system, by simultaneously executing these analyses, can gauge whether the article surpasses a predefined goodness score threshold set by us. This threshold serves as a criteria to determine whether the article is pro or against the Republican Party. Subsequently, the system can make informed

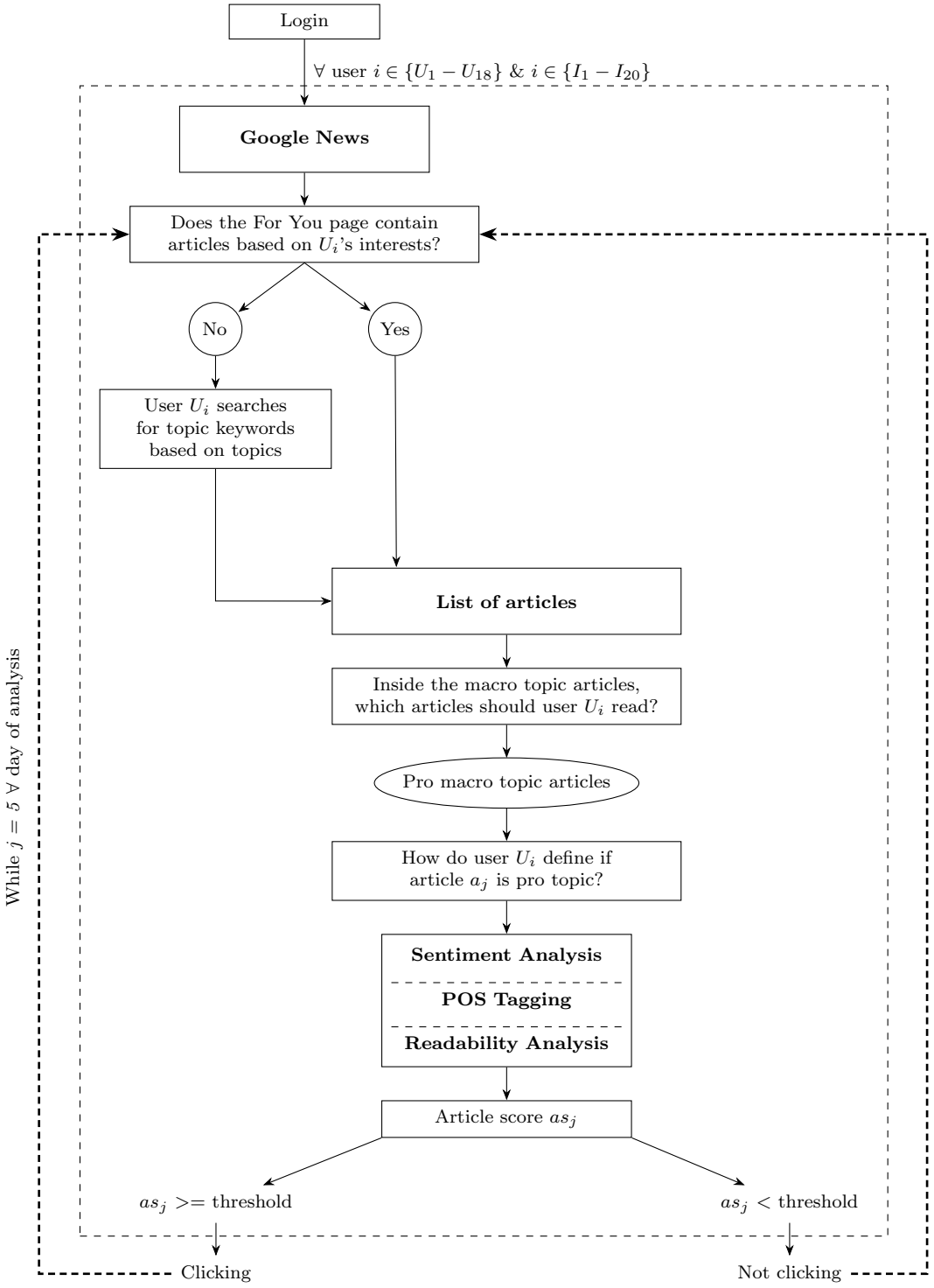


Figure 4.1: Users Interaction Flow Reading News.

decisions about clicking and simulating the reading of the article based on its alignment with the user's interests. Once the decision is made on whether to click on a particular article, maintaining a record of the clicked news becomes a straightforward process. Similar to the functionality observed in Google News, Yahoo as a news media aggregator also dynamically adjusts its content based on user interests and past news viewed. This feature simplifies the task of tracking the news articles clicked, as the platform itself tailors the content to align with the user's preferences. Taking advantage of Python's extensive library support for spreadsheet integration, we seamlessly implement a system to autonomously create and manage the dataset. This approach not only eliminate the need for manual tracking of displayed news but also removes the necessity of physically reading and simulating human behavior for each article. By leveraging Python's capabilities, we automate the process of recording and organizing the clicked news, allowing for the systematic creation of a dataset. This automated dataset generation method not only enhance the efficiency of our research but also ensure accuracy and consistency in capturing user interactions within the news media aggregator.

Table 4.2: Dataset Head containing News Articles' Information.

User ID	Majority/Minority	Topics	News ID	Title	Description	Link	Date
1	Macro	Republican Party	1	Haley heckled as Trump movement asserts its dominance at shrunken CPAC.	Republican presidential candidate Nikki Haley stepped into the hallway after speaking at the [...]	https://www.washingtonpost.com/politics/2023/03/03/haley-trump-cpac-2024-presidential/	March 04, 2023
			2	Trump proposes 10 futuristic 'Freedom Cities,' featuring Jetsons-like flying cars	Former President Donald Trump on Friday shared the broad strokes of an ambitious plan to build 10 new "Freedom Cities" [...]	https://www.cnbc.com/2023/03/03/trump-proposes-10-futuristic-freedom-cities-featuring-jetsons-like-flying-cars.html	March 04, 2023
			3	Trump calls for 'quantum leap' in standards of living through creation of 'Freedom Cities'	"Past generations of Americans pursued big dreams and daring projects that once seemed absolutely impossible," Trump said [...]	https://www.msn.com/en-us/news/politics/trump-calls-for-quantum-leap-in-standards-of-living-through-creation-of-freedom-cities/ar-AA18ctwt	March 04, 2023
	Micro	Music	1	Tour news: The Cure, Hot Chip, Black Coffee, Lisa O'Neil, The Courettes, Waste Man, more	Hot Chip will be on tour in North America this spring [...]	https://www.brooklynvegan.com/tour-news-the-cure-hot-chip-black-coffee-lisa-oneil-the-courettes-waste-man-more/	March 04, 2023
			2	HIT-BOY DELIVERS KNOCKOUT BLOW TO HITMAKA AS PRODUCER SPAT CONTINUES	Hit-Boy has continued his war of words with Hitmaka outside the booth [...]	https://hiphopdx.com/news/hit-boy-hitmaka-feud-knockout-blow	March 04, 2023
			3	March Girl Group Brand Reputation Rankings Announced	The rankings were determined through an analysis of the consumer participation [...]	https://www.soompi.com/article/1572076wpp/march-girl-group-brand-reputation-rankings-announced-5	March 04, 2023
2	Macro	Republican Party	1	Trump collaborates on song with Jan. 6 defendants	Former President Trump is featured on a new song [...]	https://thehill.com/blogs/in-the-know/3883164-trump-collaborates-on-song-with-jan-6-defendants/	March 05, 2023
			2	Jimmy Kimmel Tells Trump What'll Happen When He Finally Gets Arrested	Jimmy Kimmel made short work of Donald Trump's latest wild claim on his social media website. [...]	https://www.huffpost.com/entry/jimmy-kimmel-trump-prison$_{n6}$40182$cfe4b$0691ec73ee4f0	March 05, 2023
			3	Trump says he won't drop out of 2024 race if he's indicted	Former president Donald Trump said Saturday that he would not drop out of the 2024 presidential race if he were indicted in any of the federal and state investigations he faces.	https://edition.cnn.com/2023/03/04/politics/trump-cpac-speech/index.html	March 05, 2023
	Micro	TV	1	Silicon Valley Bank collapses after failing to raise capital	Silicon Valley Bank collapsed Friday morning after a stunning 48 hours [...]	https://www.cnn.com/2023/03/10/investing/svb-bank/index.html	March 05, 2023
			2	Kristen Doute Returning To 'Vanderpump Rules' In Season 10, Throwing Gas On The Scandal Fire	Bravo announced today that Kristen Doute, who was fired by the series in 2020 over a racist prank against former costar Faith Stowers, will be coming back in Season 10. [...]	https://deadline.com/2023/03/kristen-doute-returning-to-vanderpump-rules-season-10-1235285683/	March 05, 2023
			3	'SNL': Jenna Ortega and Fred Armisen Have Weird 'Wednesday' Reunion	SATURDAY NIGHT LIVE host Jenna Ortega and Wednesday co-star Fred Armisen teamed up again for a not-so-PG version of the 1998 film The Parent Trap [...]	https://www.rollingstone.com/tv-movies/tv-movie-news/snl-fred-armisen-plays-jenna-ortegas-crude-twin-the-parent-trap-1234694977/	March 05, 2023

Chapter 5

Insights and Data Patterns

In this chapter, we face a typology of of analysis that carefully track how the news content changes over time. Our focus is on understanding how Google News evolves, including changes in topics shown in the For You section, and how news articles are arranged. We consider the factors that affect how Google News selects and displays content each day. For each user, our initial data collection involves gathering a total of 5 news articles daily – 3 related to the designated macro topic and 2 pertaining to the micro topic, as previously outlined. While the analyses unfold over different days, their shared objective is to visualize any discernible patterns in the evolution of news content. This collective effort seeks to comprehend and interpret whether there exists a distinct development in the news, contingent upon the diverse topics. As mentioned earlier, the formation of echo chambers is influenced by how specific topics are approached. Our aim is to discern if Google News demonstrates distinct approaches and classifications for various topics. This investigation unfolds through a methodical exploration of the evolving news landscape over consecutive days, shedding light on potential patterns and variations in the treatment of different topics within the platform.

For our analysis, we leverage Python as our primary programming language for its versatility and user-friendly features. Python's ecosystem, including powerful data visualization libraries such as Matplotlib, enables us to effortlessly generate graphs and visual representations of our data.

5.1 Homophily between News Recomendation and News Consumption

Homophily is a concept that has been used to represent the tendency of individuals to form relationship with another individual who has similar interests.

Several existing research works have proposed the utilization of homophily between users to understand the formation of echo chambers in social networks. However, as previously mentioned, there is no explicit relationship among users in news media aggregators. Therefore, in this section, we propose three different forms of homophily in news recommendation on the basis of the news consumption behavior. We initially study homophily in news topic with respect to frequency, followed by more detailed temporal analysis, and finally position based analysis. Additionally, we investigate whether any news topic has higher probability to ensure similar news recommendation than the others. We also explore whether the news recommendations vary across locations with respect to the same topic.

5.1.1 Homophily in Frequency

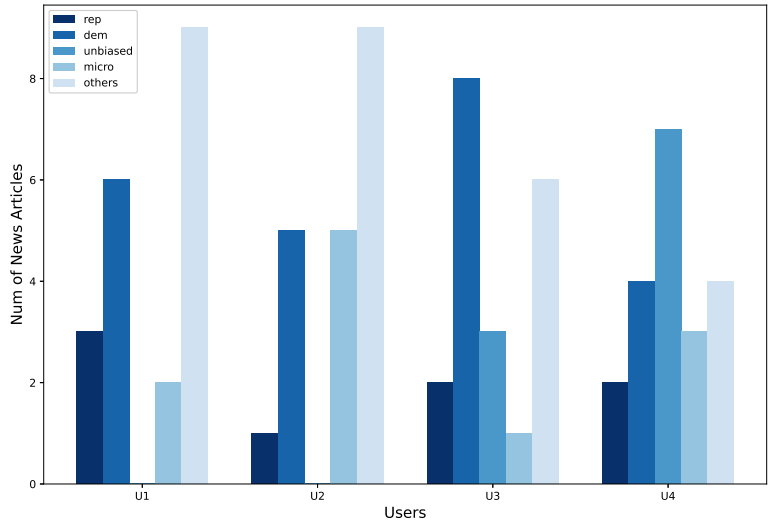
We study the distribution of news articles recommended to a user based on their news consumption behavior. We study the number of news articles that are recommended to an user on a particular day that belongs to the user's macro topic, micro topic and news which are neither macro nor micro (we refer to these as others). Usually, the nuts belong to the others section, naming the *Others* section, to political news (even for users who do not follow political topics), and news belonging to the Google News Homepage, i.e. popular and global news. The historical period in which we carried out the research has a strong impact on the number and type of news; clearly, we want to understand if this number is unbalanced if it is clearly greater than the number of news belonging to the topics followed.

We opt for a specific day and replicate this type of analysis across several days. Our focus center on the top 20 news articles featured in the For You section of each user. This approach allow us to systematically assess and compare the evolution of news content over the selected timeframe. Our observations indicate that the news recommendation varies on the basis of the topics a user follows as macro and micro topic. For example, we observe users I_1, I_2 I_{11} and I_{12} with pro-Government leaning get news related to pro-Government leaning more than the news related to Opposition leaning. We observe similar phenomenon for I_3, I_4, I_{13} and I_{14} who have pro-Opposition leaning for news related to pro-Opposition leaning. However, we do not observe the same with respect to users from USA. While users who follow democratic news, such as U_3, U_9 and U_{10} gets a higher proportion of news related to democratic rather than republican news, U_1, U_2, U_9 and U_{10} who follow Republican Party as their macro topic do not always get more number of pro-Republican news than pro-Democratic news (as showed in Figure 5.1). Additionally, the number of news articles recommended that belongs

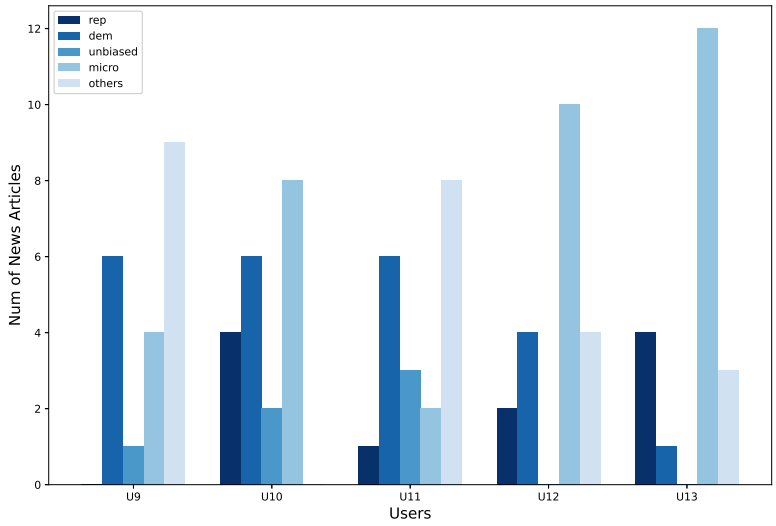
to micro news topic varies on the basis of topic of the micro news. For example, users who have Entertainment as minor news topics are recommended Entertainment news with high frequency irrespective of the users location and macro news topic. Therefore, our observations indicate that irrespective of the location and topic for macro and micro news, users generally get recommended on the basis of the topic of their choice and there are very few recommendations that do not belong to either macro or micro topic. We also observe the probability of a higher number of recommended news related to macro news topic is higher than for micro news topic.

Our observations for users with major news topic as non political news indicate that users are recommended with highest probability news specific to their macro news topic irrespective of location and the macro news topic followed by their micro news topic. We observe very few news articles are recommended which belong to neither macro or micro news topic. Additionally, we observe that news articles which cover different aspects of the news event or different news events are manuscript submitted to recommended to users on the same day irrespective of their location. This highlights that a user gets specifically tuned news article recommendation on the basis of their choice of news topic and political leaning which varies across users. To confidently discern disparities in the news provided based on the perceived "importance" or ranking embedded within Google News, we conducted a comprehensive analysis across all users, even those with no explicit interest in political news. Figure 5.2a and 5.2b provide an insightful illustration of this analysis. Notably, users U_5 and U_6 exclusively receive news from Google News pertaining to their primary topics, as depicted in the graph. The presence of these news classifications on the Google News homepage underscores their significance within the platform's content hierarchy.

Contrastingly, users U_7 and U_8 , following topics like Movies, Fashion News, AI, and Food receive a substantial number of news articles classified under the "others" category. This observation persists even after replicating the analyses over several days, indicating that Google News assigns varying degrees of importance to different topics. For instance, topics such as Food or Fashion do not receive a comparable volume of news articles compared to users engrossed in more globally significant topics like politics. This substantiates our earlier assertion and is reinforced by the infrequent appearance of news related to Food or Fashion topics on the Google News Homepage, suggesting a discernible hierarchy in the platform's treatment of diverse topics based on their perceived significance.

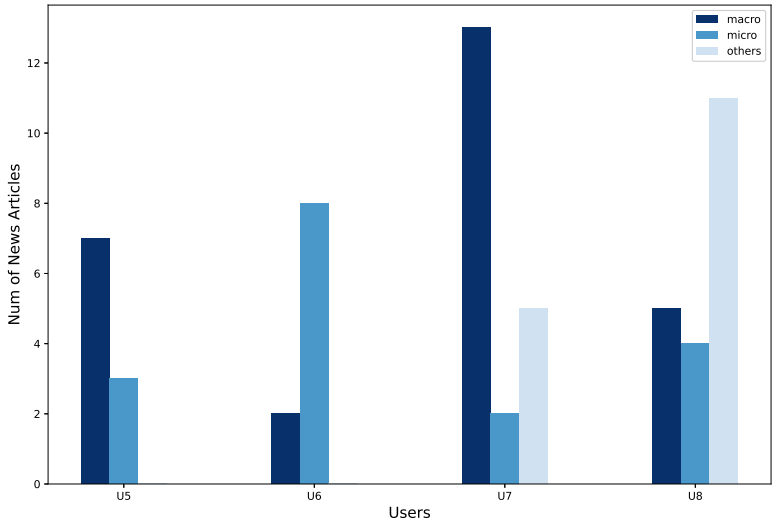


(a)

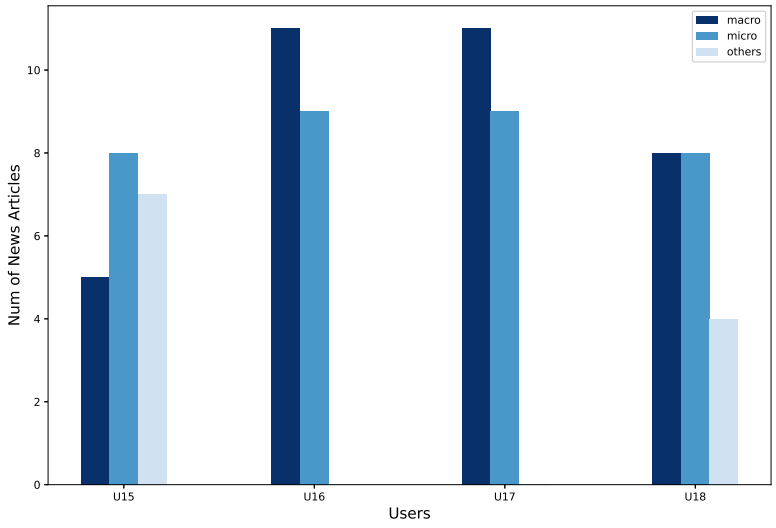


(b)

Figure 5.1: Visualizing the Ratio of News Recommended to Politic Users as Macro and belong to USA.



(a)



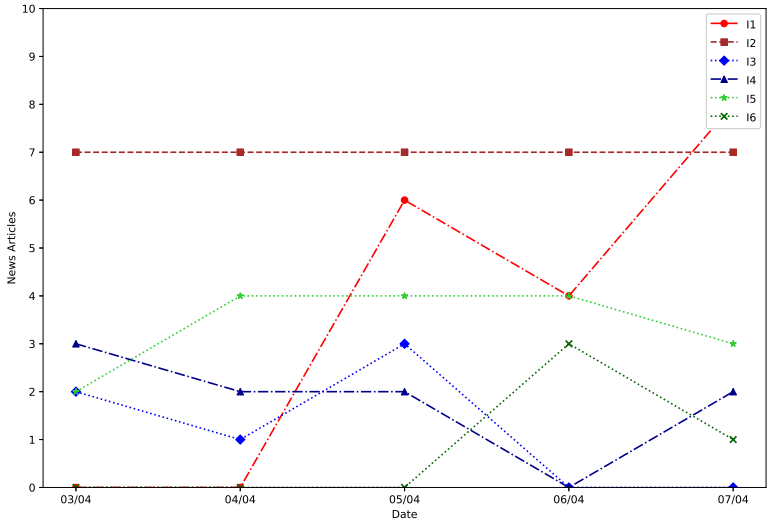
(b)

Figure 5.2: Visualizing the Ratio of News Recommended to Non-politics Users as Macro and belong to USA.

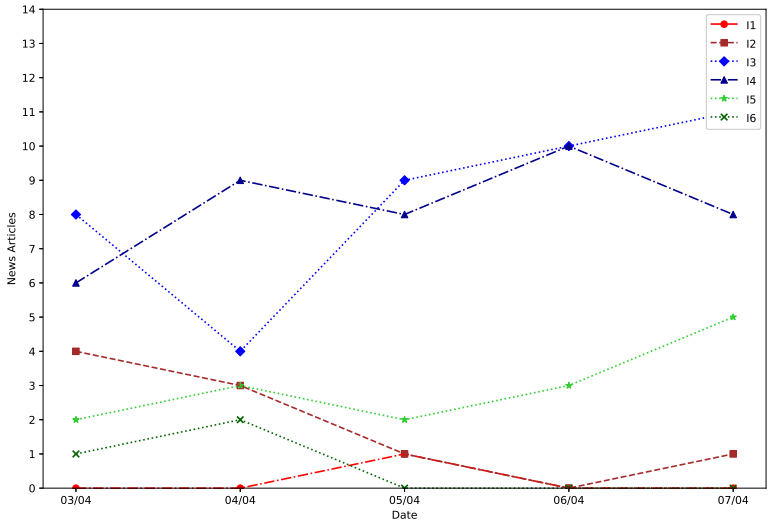
5.1.2 Temporal Homophily in News Recommendation

From our previous experiment, we can conclude that the news recommended to an user is based on the user's news topics of choice and leaning and the news recommended varies across users on the basis of the topic. Currently, we study whether the observed uniformity in news recommendation is maintained temporally. For our experiments, we consider a week and calculate frequency of news articles recommended to an user with respect to a particular news topic. For example, for the news topic politics, we calculate the number of news articles related to pro-Government recommended to an user who has politics as major news topic. We repeat this subsequently for pro-Government and pro-Opposition leaning for users from India, pro-Republican and pro-Democratic for users from USA, respectively. Our observations indicate that the number of news articles recommended to an user which belongs to her macro news topic is much higher than any other news topic irrespective of the day of the week, date or location of the user. We observe similar behavior irrespective of the macro news topic being political or non-political. We also notice, say for India, that the number of pro-Government leaning news recommended to a user with pro-Government leaning as the macro news topic is higher than the users with different macro news topic irrespective of the day, as Figure 5.3 is showing. We observe similar phenomenon for pro-Opposition leaning news.

However, for users with pro-Republican political leaning, they are either recommended news which cover republican news or pro-Democratic view point. We observe this similar pattern through different experiments for pro-Republican users in Google News, shown in Figure 5.4. A majority (at least 70%) of the news articles recommended to an user are related to a particular political leaning of their choice and it varies immensely both on the news event and framing of the same news for users with different political leaning even on the same day. We observe similar behavior for users with pro-Democratic political leaning in USA. For user with macro news topic as non political, we observe similar irrespective behavior of the macro news topic, an user gets recommended mostly the news which is their macro news topic, whether it is Sports, Entertainment, Business or Technology. We show an example in Figure 5.5.

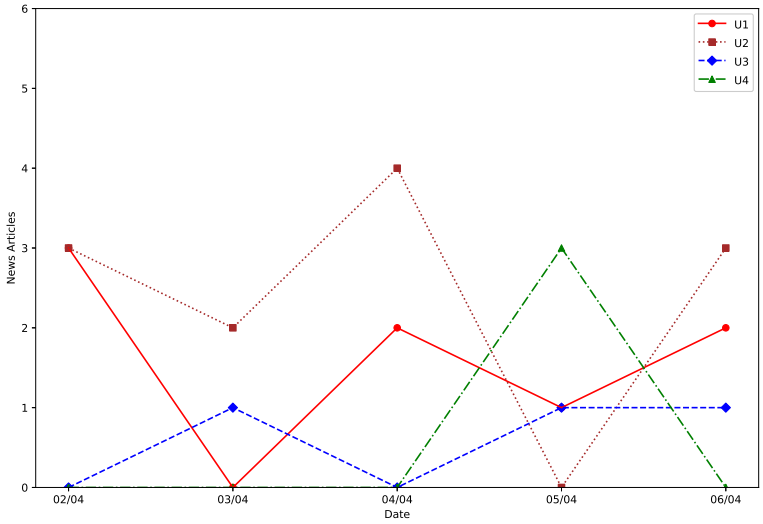


(a) Pro-Government Temporal Pattern.

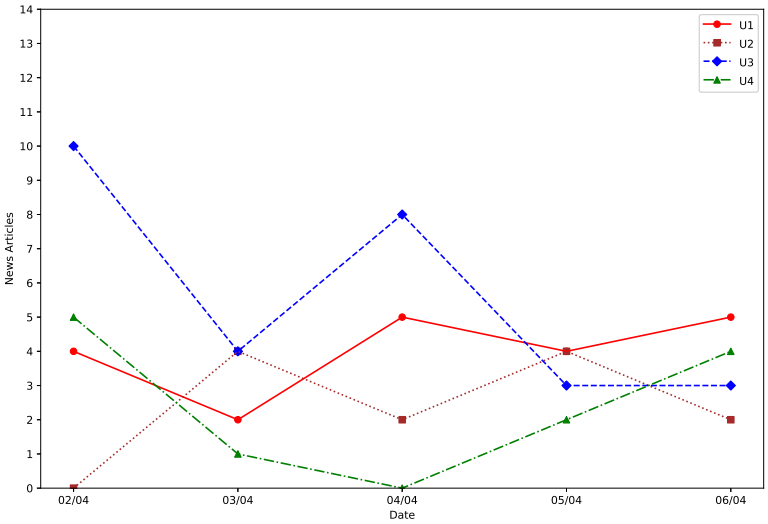


(b) Pro-Opposition Temporal Pattern.

Figure 5.3: Temporal Homophily in Political News Recommendation for India.

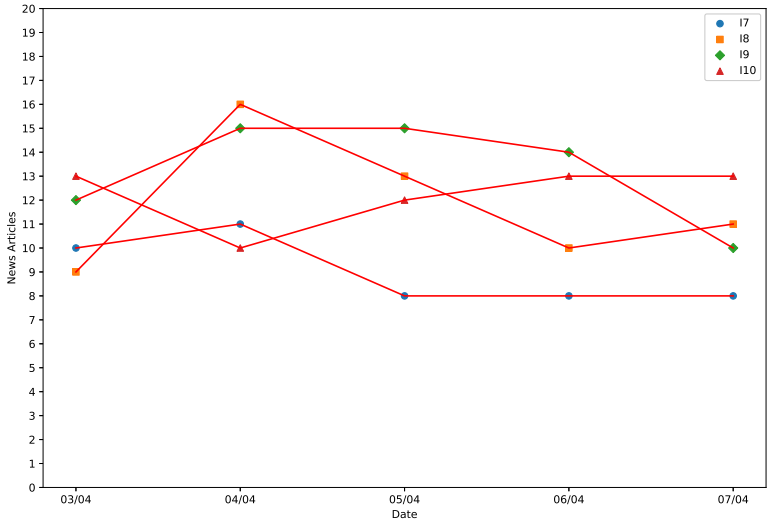


(a) Pro-Republican Temporal Pattern.

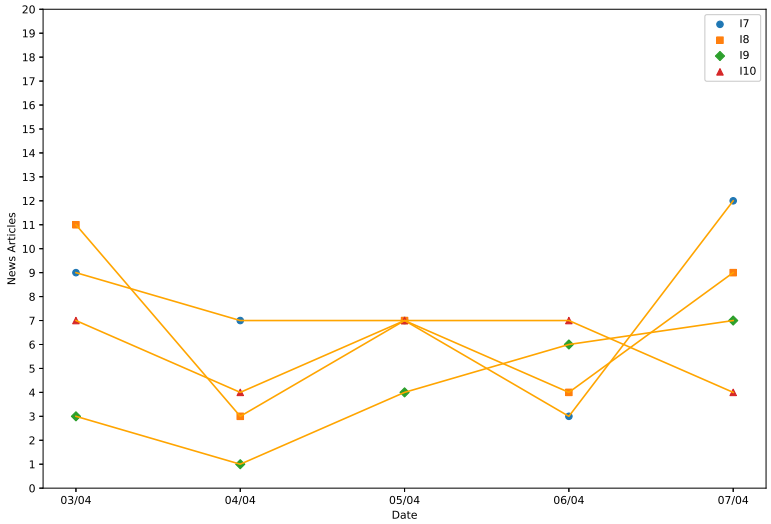


(b) Pro-Democratic Temporal Pattern.

Figure 5.4: Temporal Homophily in Political News Recommendation for USA.



(a) Macro News Temporal Pattern.

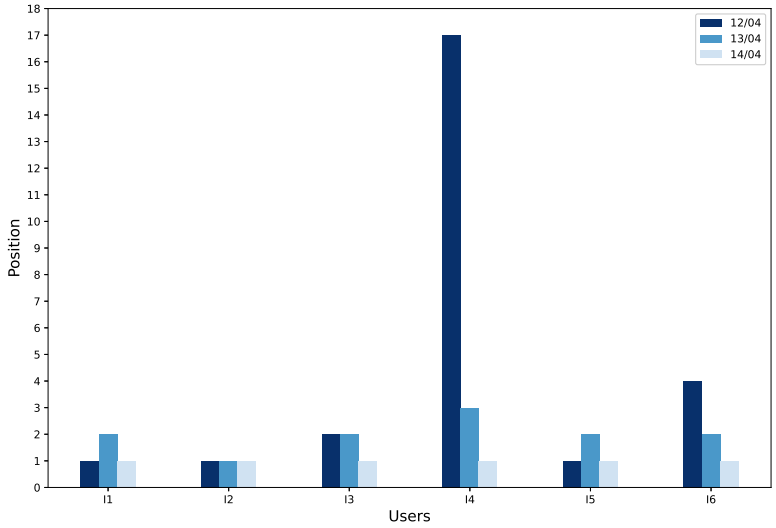


(b) Micro News Temporal Pattern.

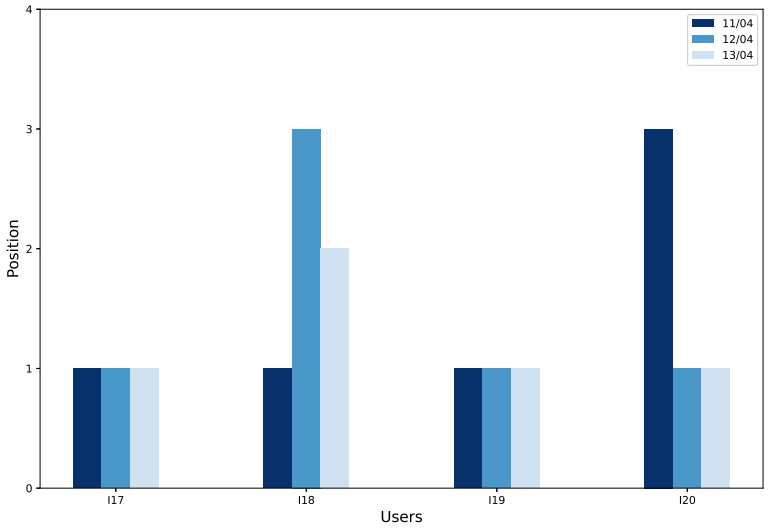
Figure 5.5: Temporal Homophily in Non-political News Recommendation for India.

5.1.3 Positional Homophily in News Recommendation

As Google News ranks the news articles based on relevance to the user, we explore the positional homophily in news recommendation for an user, i.e., we study the first position of any news article that belong to the user's macro news topic. An user has high positional homophily in news recommendation if the news articles which match her macro news topic is ranked early in her news feed. For our experiment, we repeat this for all the users of both USA and India for 5 days, respectively. Our observations indicate that irrespective of the macro news topic, the position of the first news ranges between 1-3 mostly. Additionally, we observe that for an user, the position of their macro news topic rarely appears beyond position 3 in the Google News page. This further highlights that the news recommended to different users are ranked differently on any day irrespective of the news events of that day and is dependent only on the user's news topic choice and political leaning. We further observe that it remains consistent irrespective of date and day of the week. We show few representative examples in Figure 5.6, where we show the first position of macro news topic for I_1 to I_6 in Figure 5.6a, for I_{17} to I_{20} in Figure 5.6b, respectively, for consecutive 3 days. Therefore, on the basis of this experiment, we can observe that Google News recommends and ranks news specifically aligned to every user macro news topic interests irrespective of the event of the day and the news topic.



(a)



(b)

Figure 5.6: Positional Homophily in News Recommendation.

5.2 User Similarity Analysis

Until now, our analyses have mainly focused on individual users and how their interests were influenced by the news they read. We deliberately left out any outside factors that could affect their reading habits. Now, instead of just looking at individual users, we want to compare them to see how similar they are. We consider lots of different factors to do this. This approach helps us retrieve patterns and trends that go beyond just one user. By looking at all these different factors, we hope to get a better understanding of how news changes during time. To study the variance in news recommendation among users quantitatively, we study 3 different measures, namely, *User Viewpoint Similarity Index*, *User Viewpoint Representation Index* and *User Stance Index*. Through these metrics and observations, we intend to capture how the news recommendation is similar between a pair of users on the basis of their macro, micro news topic and political leaning. We follow for the definition of User Viewpoint Similarity Index and User Viewpoint Representation Index. This allows us to have different indices, to be able to compare the similarities and differences from users with dissimilar and similar interests from every perspective.

5.2.1 User Viewpoint Similarity Index

The first index we address in this section is called *User Viewpoint Similarity Index*. For this type of analysis, we compare similarity in the news recommended between a pair of users on the basis of the topics of the news.

We consider the topics as a combination of all the possible macro and micro news topics. Suppose we approach the situation from the perspective of users from India; the first example we report concerns the situation of the topics according to the point of view of users from India, i.e., pro-Government, pro-Opposition, Neutral, Sports, Entertainment, Technology, World and Business, for a total of 8 topics. Therefore, for an user I_j, *Topic Distribution* is a vector of size 8 where each position of the vector represents each topic, and the value is the frequency of the news articles recommended to I_j on that topic. Suppose we take two users: user I_1 which is pro-Government and with interests to Sports news, and user I_2 which is also pro-Government and interested on Entertainment news. Our earlier checks tell us that Sports and Politics news usually get a good spot in the ranking. That's a hint that Google News thinks these topics matter on a larger scale. Now, following our plan, we'd expect a topic distribution vector with high news frequency at the top (for pro-Government, let's say position 1) and another spike further down the line (let's call that Sport at position 4). Same way for user

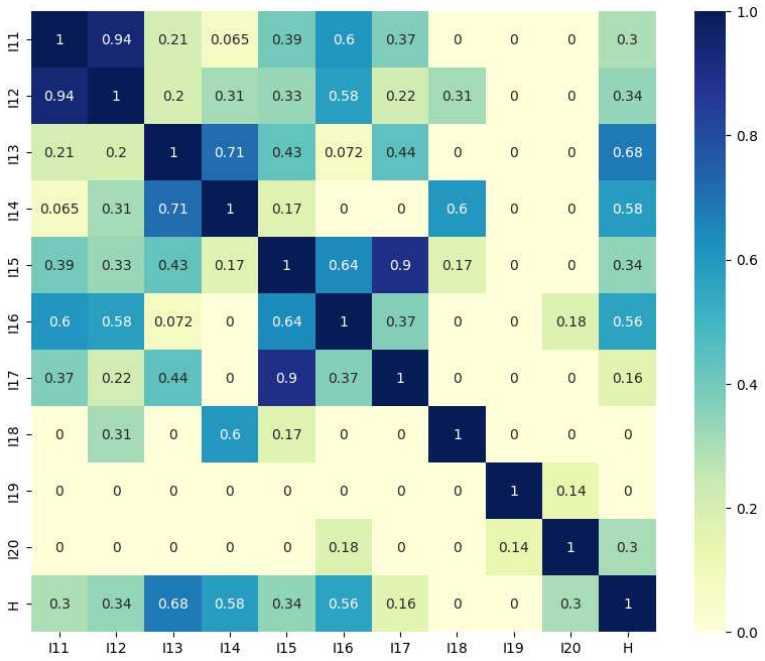


Figure 5.7: User Viewpoint Similarity Index Matrix.

I_2.

We calculate *User Viewpoint Similarity Index* between two users I_1 and I_2, as the weighted cosine similarity[1] between Topic Distribution I_1 and Topic Distribution I_2. We repeat this for all pair of users to construct the complete matrix. Our observations as shown in Figure 5.7 indicate that User Viewpoint Similarity Index is higher if the users have same macro news choices, such as, I_1 and I_2 have a similarity score of 0.43. We can also see that user I_3 and I_4 have a similarity score of 0.38, whereas it varies significantly if the users like political news but with different stance. Additionally, we observe that Entertainment is a very popular news topic worldwide and has huge number of recommended news which affects User Viewpoint Similarity Index. Analyzing user similarity through a similarity matrix allows us to understand the personalized content recommendations based on user preferences and interests, fostering user engagement and satisfaction. By examining topic distribution vectors and similarity indices, insights into user behavior can be gained, helping to understand specific topics that resonate with distinct

[1]Variant of cosine similarity that incorporates weights for the dimensions (features) of the vectors being compared. It's a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them.

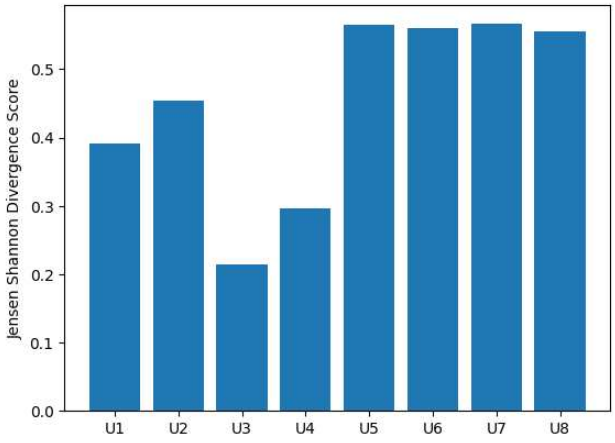
user groups. This information can be leveraged to improve content curation and prioritize the display of news articles aligned with users' preferences. Furthermore, the similarity matrix aids in identifying and mitigating group of people with similar interests where users are exposed to a narrow range of perspectives reinforcing their existing beliefs. Clusters of users with similar viewpoints can be detected, and measures can be implemented to expose these groups to diverse perspectives, fostering a more open and balanced information environment. It also contributes to algorithmic fairness, ensuring unbiased content recommendations and promoting a diverse user experience.

5.2.2 User Viewpoint Representativeness Score

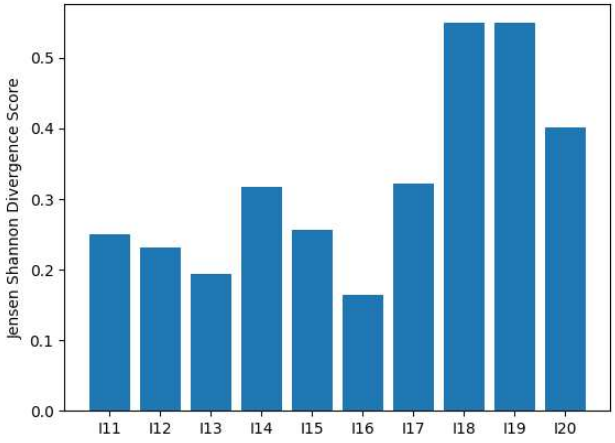
We use *User Viewpoint Representativeness Score* to understand the level of specificity in news recommendation for an user, and how it varies than the news articles published that day. Although Hada et al. [Hada et al., 2023] propose User Viewpoint Representativeness Score, their proposed metric captures the difference in information visible amongs users in a social media platform and therefore, is not applicable for news recommendation in news media aggregator. Based on the intuition of User Viewpoint Representativeness Score, we propose this variant to capture the variance in news recommended specifically to an user with respect to all the news articles published on that day.

The distinction lies in recognizing that news media aggregators operate in a domain where content curation is not only influenced by user interactions but also by the broader landscape of news topics. Unlike social media platforms, where user-generated content predominantly shapes the information visible to users, news media aggregators incorporate a diverse range of news articles from various sources. Our variant of the User Viewpoint Representativeness Score takes into account this unique characteristic by focusing on the specific news articles recommended to an individual user. This metric becomes particularly relevant in assessing the level of specificity in news recommendations – providing insights into how tailored the suggested content is to an individual user's preferences and viewpoints, especially when compared to the broader spectrum of news articles published on the same day.

We calculate User Viewpoint Representativeness Index as the *KL-divergence* between Topic Distribution (I_1) and Topic Distribution (I_H), where Topic Distribution (I_1) and Topic Distribution (I_H) represents the distribution of the news articles with respect to different topics for I_1 and news published that day, respectively. The Kullback-Leibler (KL) divergence, also known as relative entropy, is a measure of how one probability distribution diverges from a second, expected probability distribution. In the context of calcu-



(a)



(b)

Figure 5.8: User Viewpoint Representativeness Score for USA and India, respectively.

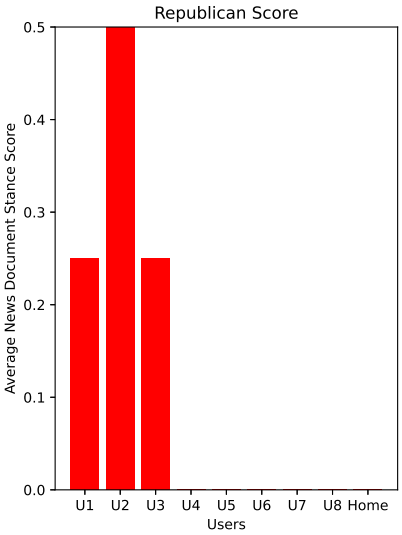
lating the User Viewpoint Representativeness Index using KL-divergence, it provides a numerical measure of the information difference between the topic distribution of a user (I_1) and the overall distribution of news articles (I_H). It quantifies how the user's preferred topics deviate from the general distribution, helping to assess the specificity and uniqueness of the user's viewpoint. By calculating this score, we can identify the topics that significantly contribute to the difference between the user's viewpoint and the overall distribution. This information is valuable for understanding which topics are particularly relevant to the user and contribute the most to the representativeness index. We repeat this experiment for all the 38 users daily for 7 days.

Therefore, a high KL divergence score of I_1 signifies a low User Viewpoint Representativeness Score, i.e, I_1 receives very few news among all the news published on that day. Our observations indicate that the most of the users have very low User Viewpoint Representativeness Score score, i.e., very high KL divergence score around 0.4 to 0.7. We show a representative example in Figure 5.8: it shows of User Viewpoint Representativeness Score for U_1 to U_8 (Figure 5.8a), and for I_{11} to I_{20} (Figure 5.8b). The Jensen-Shannon Divergence (JSD) represented on the graph is a measure of similarity between two probability distributions. It is derived from the Kullback-Leibler (KL): it is symmetric, and this property ensures that the User Viewpoint Representativeness Score considers both the user's viewpoint (P) and the overall distribution of news articles (Q) equally. It quantifies the dissimilarity between the user's preferred topics and the general distribution of news articles. A higher JSD implies a more unique or specific viewpoint, while a lower JSD suggests alignment with broader trends. Therefore, our observations indicate that the news recommended to an user varies by a high margin with the news published on that day irrespective of the macro and micro news topic of the user, day and her location. Therefore, this highlights the low coverage of the news articles published on a day which are recommended to an user.

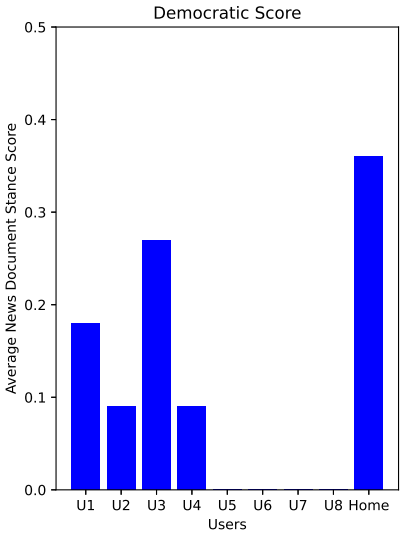
5.2.3 User Stance Index

For user political leaning based Viewpoint Analysis, we study the fraction of news articles recommended to an user for a particular political leaning on a day. This provides us with an understanding of variance in the recommended news across different users based on their news preferences and political leaning. For example, on any particular day, we calculate the number of news articles recommended to an user with respect to a particular political leaning. Therefore, suppose, for political leaning as Republican, we calculate the number of news articles with Republican political leaning recommended to an user. Our observations indicate that users are recommended news that match with their preferred political leaning irrespective of location. Additionally, we observe that Google News recommends democratic news and neutral political news rather than republican news to an user with no political leaning or news reading behavior. For example, we observe in Figure 5.9a, U_1 and U_2 are users with Republican political leaning, so the republican political news are mostly recommended to them followed by U_3 and there is no republican news recommended to an user with Neutral leaning, U_4 and no political leaning, such as, Home [2]. However, as shown in Figure 5.9, Google

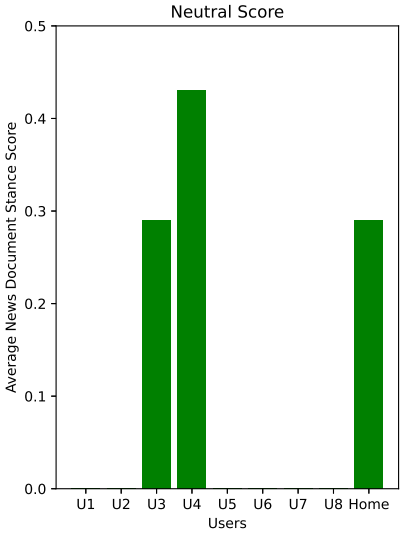
[2]We refer to the Google News' Homepage, as we did in the previous sections.



(a)



(b)



(c)

Figure 5.9: User Stance Index for USA.

News recommends democratic news to recommended to an user with Neutral leaning, U4 and no political leaning, such as, Home.

We can observe in Figure 5.9a a notable political stance score favoring the Republican Party for users U_1 , U_2 , and U_3 —indicating a political inclination towards the Republican Party among these users. As highlighted before, an interesting observation emerges from the Home section of Google News,

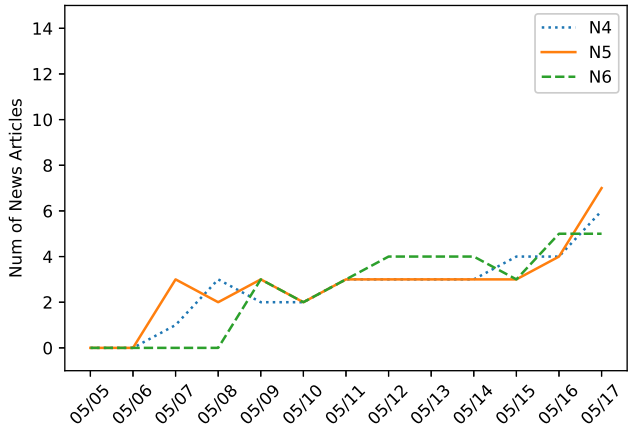
which predominantly features important global news, including a significant portion related to politics. Notably, this section tends to showcase primarily Democratic or unbiased political news. This trend is consistent across different user inclinations. For instance, Democratic and unbiased users predominantly encounter news aligned with their political stance, but Republican-oriented news is notably absent. This reaffirms our earlier findings, suggesting a one-sided presentation of news. Specifically, pro-Democratic news is consistently delivered to Republican users, while the reverse—presentation of pro-Republican news to Democratic users—is not observed in a comparable manner. This asymmetry in news representation underscores the potential for selective exposure and bias in the information presented to users with specific political inclinations on Google News.

5.3 New Users in Echo Chambers

For a new user, Google News Recommender provides news based on the popularity of the news event with respect to the other news events. As observed, the quantity and selection of news presented to users are influenced by various external factors. For instance, during a hypothetical election period for a political party, there may be a surge in the number of news articles related to politics compared to other topics. External events, seasons, or specific periods can significantly impact the news landscape, leading to variations in the content presented to users.

However, with implicit feedback from the users on the basis of the news they read regularly, Google News specifically suggests news articles for an user. Our experiments and observations from previous sections shows that the news recommended to an user depends on the users's macro and micro news topic. Additionally, users with similar topical interest and political leaning are shown similar news which might be drastically different from the other users. Therefore, in order to understand this timeline of generic news recommendation to specifically tuned news recommendation, we perform an experiment where we create a batch of 6 new users (N_1 to N_6), of different news topic choices, and simulate how their news recommendation behavior changes daily. For example, for a batch of six users, we consider the macro news topic of N_1 as Entertainment, N_2 as Sports, N_3 as Technology, N_4 as pro-Government leaning, N_5 as pro-Opposition leaning and N_6 has Neutral political leaning. We create 4 batches of 6 users each such that 2 batches are for users who belong to India and the other two are for users who belong to USA. We observe how the number of news articles recommended with respect to their news choice changes every day while considering that every

(a)



(b)

Figure 5.10: New Users' News Temporal Evolution.

user reads 5 news articles daily which are related to their specific news choice. Our observations indicate that it takes more number of days for an user with specific political leaning to get a majority of its news feed aligned to that specific political leaning than an user who reads Entertainment or Sports based news. This is logically attributed to the need for a sufficiently large number of news articles. The algorithm must discern not only the user's interest in political topics (which is relatively straightforward to differentiate from other topics) but also determine the specific political party alignment of the user. In contrast, users interested in Sports or Entertainment-related news find their preferences recognized more swiftly, as these topics are more

easily distinguishable. Subsequently, the following sections will delve into the complexities of identifying a user's political leaning, given the multifaceted interpretations that concepts within an article can have from different perspectives, especially when processed automatically by algorithms. Additionally, we observe that an user with macro news topic as Sports or Entertainment gets only news related to Sports and Entertainment. This observation aligns with our previous analyses, confirming that subjects like Sports and Entertainment fall within the category of topics classified by Google News as important and popular. In an hypothetical ranking, Sports and Entertainment would likely secure top positions. Consequently, Google News doesn't feel the need to diversify the user experience by presenting recommendations on other news types. The reasoning behind this is that these important topics already encompass a significant portion of the user's interest space. We show representative examples for N_1 to N_3 in Figure 5.10a and N_4 to N_6 in 5.10b, respectively. Therefore, through this study, we explore how the Google News Recommender gets tuned to the specific user interests based on the reading behavior of the user irrespective of the news published that day and recommended to users who have no specific interest. We also observe that the number of days to get specifically tuned depends on the topic of the user and the number of news articles she reads. This study also provides us an intuition about the threshold number of days required for an user to be in an echo chamber.

5.4 Case Study: Susceptibility to Propaganda News

In this section, we delve into examining whether there exists a correlation between the user's chosen news topics and their susceptibility to being recommended news with propaganda. Propaganda, in the context of news articles, refers to the dissemination of biased or misleading information intended to shape public opinion or promote a particular agenda. It often involves the use of persuasive techniques to influence individuals' beliefs or attitudes, rather than presenting objective and unbiased reporting. Propaganda in this context can take various forms, such as selective story choices, framing, or biased language, aimed at influencing public opinion or promoting a particular agenda. Aggregators may choose to highlight or suppress certain stories based on their alignment with a particular political, social, or economic perspective. This selective curation can shape the narrative presented to users. Or, the way news stories are framed can influence how readers perceive events.

Aggregators may use language that subtly guides the audience towards a specific interpretation or viewpoint, contributing to a biased understanding of the news. Manipulative headlines can be used to grab attention or provoke a specific emotional response. Also, some news aggregators personalize content based on user preferences, creating an echo chamber where users are exposed primarily to information that aligns with their existing beliefs. This can reinforce pre-existing biases and limit exposure to diverse perspectives.

Our analysis aims to uncover whether users who engage with specific news topics, especially those related to politics, might encounter a higher likelihood of receiving news articles with propagandistic elements. This investigation is crucial in understanding how users' topic preferences may expose them to potentially biased information and contribute to the formation of echo chambers where individuals are consistently presented with content that aligns with their existing beliefs. Additionally, we study if there is a relationship between the user topic choices and her susceptibility to being recommended news with propaganda, and if it varies across users on the basis of their macros and micro news topic. For this experiment, we initially observe the number of news articles recommended to a user which has propaganda. We repeat this experiment for all the users. We follow Morio et al. [Morio et al., 2020] to detect whether a news article has propaganda or not. Morio et al. [Morio et al., 2020] proposes a transformer based model which utilizes the sentence embedding coupled with named entity and pos embedding to detect whether a sentence has propaganda or not and consider a news article has propaganda if any of the sentences is propagandastic in nature. Our observations indicate that users with macro news topic with a specific political leaning has higher likelihood to be recommended news which has propaganda than the other users irrespective of the day. For example, as shown in Figure 5.11, I_1 to I_4 are recommended $2-4$ news articles which has propaganda in comparison to I_8 to I_{10} irrespective of the political leaning and the day. Since, we observe that political news has higher likelihood to have propaganda and furthermore, our previous experiments confirm that the users are recommended news on the basis of their macro news topic, we can conclude that users with macro news topic with a specific politics leaning has higher likelihood to be recommended propaganda news. This study further indicates that this can be of huge concern and requires users to be aware of their news feed and susceptibility. It also highlights the requirement to develop news recommender approaches for Google News which specifically ensures prevention of propaganda based news recommendation to users and specifically, to users with a political leaning as they are more susceptible than their counterparts.

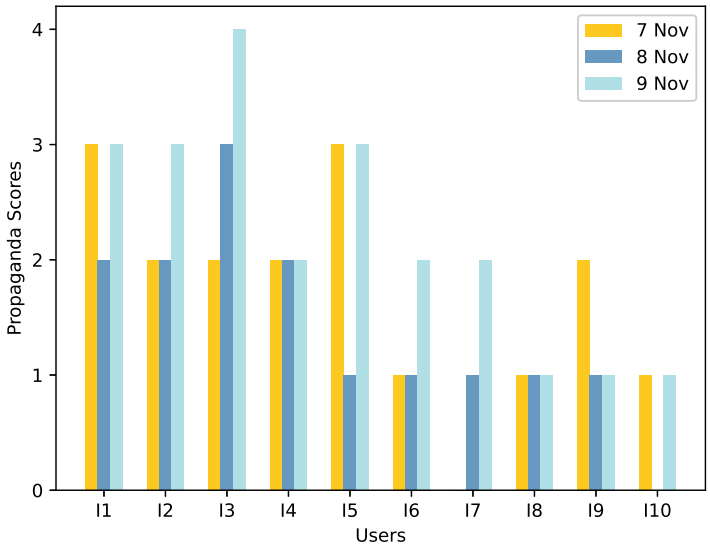


Figure 5.11: News Recommendation with respect to Macro Topic for Users from India.

Chapter 6

Content Analysis and Retrieval

Until now, our focus has been on examining how different users, with diverse interests, engage with news over time. We explored the evolutionary trend of news, starting from an empty pool of news and gradually building a dataset containing a collection of news for each user. Each user, on a daily basis, interacted with different news, altering the trajectory of the news they were exposed to. In our previous analyses, the emphasis was on what Google News offered users in terms of both the quality and quantity of news based on specific topics. We observed distinct approaches for users with specific political inclinations compared to those with entirely different interests, such as topics featured on the Google News Homepage like Sports or Global news outside of Politics. In addition to tracking the evolution of news, we delved into comparing how Google News diversified content for different users, whether they shared similar or varied interests, studying the evolution of news from various perspectives.

In this section, our focus shifts to the actual content of the news presented to each user. By actual content, we mean a detailed analysis of the article text for each news article proposed to the user, employing various text analysis techniques. Specifically, we can examine the tone used within the article, analyse the use of words and adjectives, dissect individual components of sentences, and scrutinize the overall composition of sentences. Through these analyses, we generate different scores to determine whether an article is inclined towards or against a particular topic it addresses.

6.1 Text Extraction Approach

Before delving into the analyses conducted, we need to focus our attention on a preliminary discussion concerning the panoramic vision of the process we have meticulously followed. This comprehensive task starts with the retrieval of physical links of news articles from the dataset, followed by the extraction of the textual content embedded within these links. Central to our methodology is the exploration of different approaches to text preprocessing. Prior to directly examining the textual content, preprocessing serves as a pivotal step. It involves the systematic "cleaning" of the text article, retaining only the essential components that contribute significantly to a correct analysis of the text. These components primarily include keywords and phrases that bear substantive relevance. The overarching goal of preprocessing is to eliminate superfluous elements, ensuring that the subsequent analysis is focused, accurate, and meaningful. In the ensuing subsections, we delve into a detailed exploration of the diverse facets of text preprocessing. Each aspect, from tokenization and part-of-speech tagging to named entity recognition and Sentiment Analysis, plays a crucial role in refining the raw textual data. In essence, our methodology is an intricate interplay of data retrieval, text extraction, and sophisticated preprocessing techniques. This thorough process ensures that the subsequent analyses and results are grounded in a well-processed, representative dataset, providing users with a comprehensive and meaningful exploration of the news articles under consideration.

6.1.1 Extracting Urls from Dataset

As mentioned earlier, our primary objective is understanding where we got these article links, and more importantly, why we need them. In essence, we create a dataset containing news repository clicks along with their respective links. These links serve as the entry points to the actual news articles. Automated link extraction streamlines the process, ensuring efficiency and simplicity. However, during analysis, we occasionally deviated from the original dataset. For instance, consider an analysis that spans a specific period, let's say ten days. In such cases, instead of continuously referring back to the original dataset, we took a shortcut. We created a subset of text articles, noting only the article links within that specified time frame. This approach stems from the nature of certain analyses, particularly Text Analytics, which was one of our main focuses. This analysis entails a more thorough examination of the article content. To become familiar with the news and understand the specifics for a particular user, we occasionally clicked and read the articles manually. This personalized approach allowed us to gather insights beyond

what was available in the initial dataset. The process, though involving some extra steps, contributed to a more user-focused Text Analytics analysis.

In the process of preparing the textual content for subsequent analyses, our initial step involves the extraction and preprocessing of the article text. This intricate task is accomplished through the implementation of a Python script, specifically designed to leverage the capabilities of the *Newspaper* combined with *BeautifulSoup* libraries. The Newspaper library, tailored for web scraping, plays a central role in efficiently extracting and parsing articles from various news websites. This library provides a comprehensive interface, facilitating the retrieval of key information such as the article's title, text, author, publication date, and more. BeautifulSoup is used for web scraping purposes to pull the data out of HTML and XML files. It provides idioms for iterating, searching, and modifying the parse tree. We take advantage of the library to delete superfluous parts of the text, i.e. the text strings of the advertisements (usually positioned at the sides of the article), the tags of a particular site (usually reported at the bottom of the text article), the authors, etc. We start the extraction process by providing the link to the target article, ultimately obtaining a singular string encapsulating the entire article's content.

The libraries mentioned above are not the only ones helpful for this preliminary phase. By copying links from Google News, being a news aggregator, before actually entering the news site, the user first enters the aggregator site, which consequently takes the user to the original news site. Let's take an example: *https://news.google.com/articles/CBMiTGh0dHBzOi8vd3d3Lm55dGltZXMuY29tLzIwMjMvMDUvMDIvb3Bpbmlvbi9yb24tZGVzYW50aXMtd29rZS1taW5kLXZpcnVzLmh0bWzSAQA?hl=en-US&gl=US&ceid=US%3Aen* this string contains the link to the site where the original news is located in the Google News repository. Clearly, it has its own repository, so that it can aggregate news, and in this way it can better understand the user's interests. And this hyperlink refers to the link that leads to the real news. Once the user clicks on the link to the Google News repository, containing the original news, they are automatically forwarded to the news site. This was a problem for the original link extraction operation, since, by giving only the link seen before as input, it was impossible for the system to automatically access the original article. For this, we used the *requests* library. The requests library is a popular Python library for making HTTP requests. It simplifies the process of sending HTTP requests and handling the associated responses. With requests, we can interact with web services, retrieve data from URLs, and perform various HTTP operations. With a get request, passing the Google News article as a parameter, we were able to obtain the original URL via a get request. In this way, the link

from the google news one is encoded in the standard string, for example, *https://www.nytimes.com/2023/05/02/opinion/ron-desantis-woke-mind-virus.html*. In this example, it's an article from Nytimes, a famous New York newspaper that covers political news. Once the real link was obtained, it was possible to apply the real extraction of the article text.

The extraction process encompassed retrieving essential components of the article, including the *title*, *subtitle* (if present), *description* (if present) and the *text* of the article. The description, in this context, refers to the paragraph typically positioned after the subtitle but before the main text. This paragraph functions as an abstract or summary of the text, providing a concise overview of the article's content. Furthermore, the script extracts the main body of the article. This comprehensive approach to text extraction not only enables an understanding of the article's content but also lays a solid foundation for the analysis. Through this meticulous process, we successfully obtain a complete representation of the news article text.

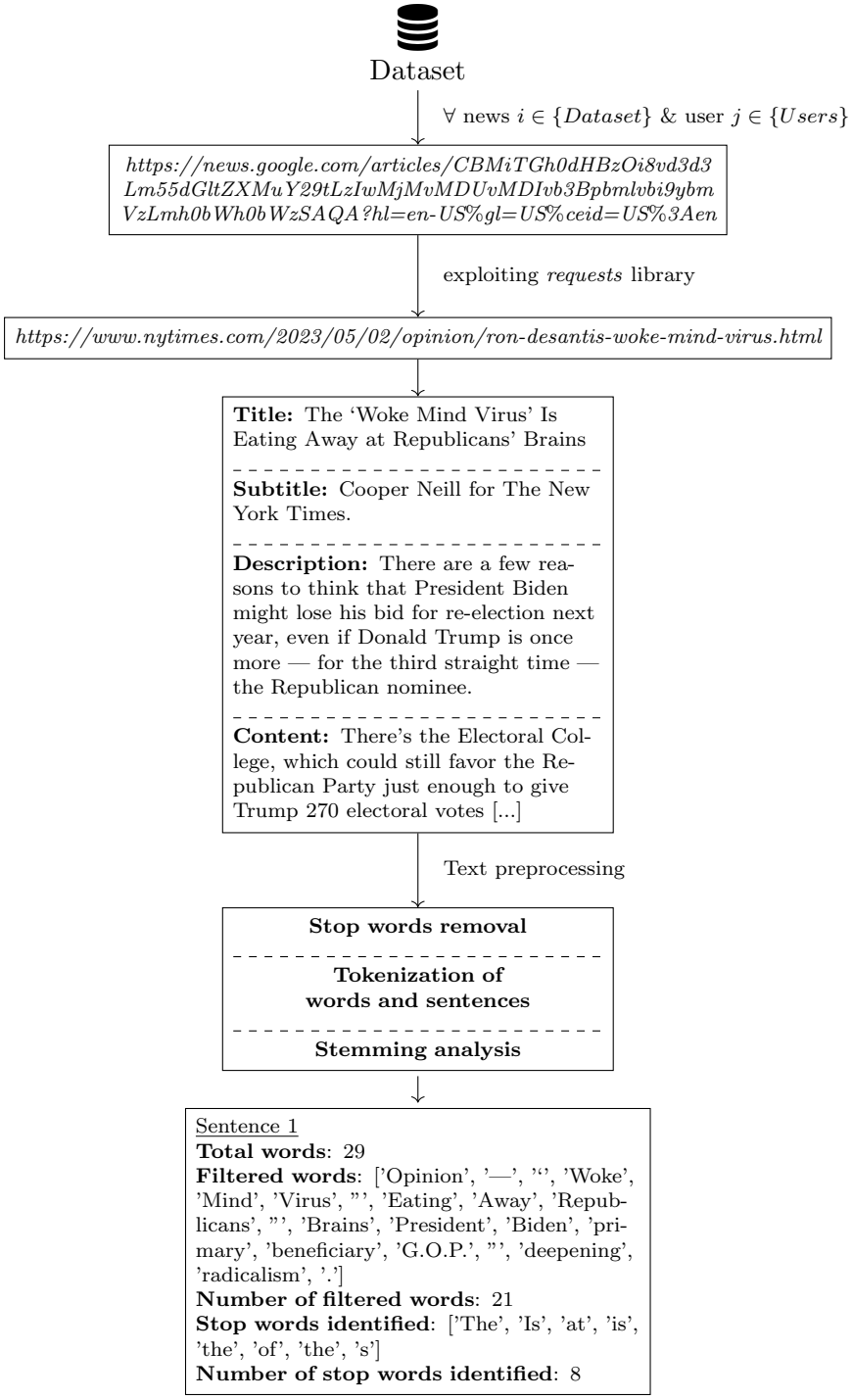


Figure 6.1: Text Article Extraction Flow.

6.2 Text Pre-processing Overview

In the previous section, our emphasis was on acquiring the essential components of a news article, which were then concatenated into a coherent textual string. The automated extraction of the article's textual content from the news link was carried out to the elements of the article, deliberately excluding words or text originating from advertising sources typically found in the peripheral sections of websites. It is important to note that these components, representing the core content of the article, are usually extractable through HTML or XML objects distinct from those encompassing the main textual body. However, due to the inherent variability in the structures of web pages, not all adhere to a standardized format. Instances exist where pages choose to integrate advertising text within subject lines that differ from those adopted by other news aggregation sites. As a result, a methodical and exacting analysis was undertaken to ensure the extraction of a refined string, containing the indispensable information required for a comprehensive article analysis.

Let's now address the topic of Text Pre-processing: it involves a series of analyses aimed at refining raw text data, transforming it from unstructured text strings into analyzable objects. This process begins with cleaning the text, which entails removing irrelevant elements such as punctuation, special characters, and stopwords. After cleaning, the text undergoes tokenization, breaking it down into individual words or tokens. Following tokenization, techniques like stemming or lemmatization may be applied to reduce words to their base or root form, thereby consolidating variations of words and enhancing analysis accuracy. The final object is a processed text that contains only meaningful keywords relevant to the topic discussed in the articles, facilitating deeper analysis and insight extraction.

6.2.1 Tokenization of Sentences

Tokenization is a fundamental pre-processing step in Natural Language Processing (NLP) that involves breaking down a text into smaller units, typically words, phrases, or symbols. In the context of sentence tokenization, the process involves segmenting a text document into individual sentences based on certain rules or patterns. Sentence tokenization is important in NLP because many NLP tasks, such as Sentiment Analysis, machine translation, and text summarization, require input text to be split into sentences so that the task can be performed on a sentence-by-sentence basis.

Tokenization of sentences is a crucial step in text preprocessing that involves breaking down a continuous string of text into individual sentences

or tokens. During tokenization, a text document is scanned, and boundaries between sentences are identified based on punctuation marks like periods, exclamation marks, or question marks. Each identified sentence is then treated as a separate unit or token, enabling subsequent analysis to be performed on a sentence-by-sentence basis. This type of tokenization, combined with that of individual words, allows the single block of text to be broken into smaller units. This phase is crucial for our analysis, as it allows us to analyse sentence by sentence, each with its context, and furthermore to analyse the individual words that make up the sentence. The library that provides us with the automated tools, and allow us to exploit them is called Natural Language Toolkit (NLTK) for Python. It's a powerful library designed to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources, such as WordNet. Additionally, it includes a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and more. NLTK is widely used in NLP and computational linguistics for tasks like text analysis, Sentiment Analysis, and machine learning applications involving textual data. Its versatility and extensive functionalities make it a popular choice for researchers, developers, and students working on projects related to natural language understanding and processing. Once this technique has been used to break the block into sentences, we obtained an object (in the specific case, a dictionary, or a structured object that allows the objects within it to be indexed) which encapsulates the different sentences that make up the text article.

6.2.2 Tokenization of Words and Stop Words Removal

Once sentences are tokenized, the subsequent tokenization of words dissects each sentence into its constituent words, where each word becomes a separate token. This process serves various purposes, including facilitating text analysis, aiding in text preprocessing for tasks like Sentiment Analysis and topic modeling.

Once we split the different sentences, and within the sentences obtained the individual tokens corresponding to each word and punctuation mark, we continue by applying a stop word removal operation. It consists in a crucial step that involves the elimination of common words, known as stop words, from a given text. Stop words, such as "the," "and," and "of," are frequently occurring words in a language that generally do not contribute significant meaning to the context of a document. The purpose of removing stop words is to enhance the quality and relevance of the analysis by eliminating terms that are considered noise. By discarding stop words, the analysis focuses on the more meaningful and content-rich words, allowing for a more accurate

understanding of the text's underlying themes, sentiment, and key topics. For each sentence, we identify the *number of total words*, the *filtered words* (the remaining keywords after stop words removal), the *number and the identified stop words*. This operation gives the user an idea of how a text article is composed.

6.2.3 Stemming Analysis and Lemmatization

The following step is to reduce words to their root or base form, known as the *stem*. This techniques is called Stemming Analysis, and its goal is to simplify word variations and treat different inflections or derivations of a word as a common representation. Stemming involves removing prefixes or suffixes from words to obtain the root form, even if the resulting stem may not be a valid word. For instance, consider the words "run," "running," and "ran." The stem for all these words after stemming would be "run." Stemming helps in standardizing words and reducing them to a common root, making it easier to analyse and process text data. It's important to note that stemming might result in the stems not being actual words, as the process is rule-based and focuses on removing affixes rather than considering the linguistic context. While stemming is more aggressive in reducing words, it is computationally less intensive compared to lemmatization, another technique for word normalization. In order to calculate a score based on the use of words in a certain context, it is essential to use Stemming Analysis as preprocessing of the article.

Unlike stemming, which involves removing suffixes from words to obtain a root form (sometimes resulting in non-real words), *lemmatization* considers the context of the word and aims to transform it into a valid word lemma. Lemmatization is a technique that involves reducing words to their base or root form, known as the lemma. The purpose of lemmatization is to normalize words so that different inflected forms or variations of a word are treated as a single, common representation. News articles often contain variations of words due to different tenses, plural forms, or derivations. Lemmatization and stemming help reduce the vocabulary size by consolidating these variations into their base forms, making it easier to analyse and interpret the text.

6.3 Text Analytics Techniques

Natural Language Processing (NLP) is a facet of artificial intelligence dedicated to facilitating communication between computers and humans through

natural language. The overarching objective of NLP is to empower machines with the ability to comprehend, interpret, and generate human-like text in a meaningful and valuable manner. Its applications span various tasks, including text understanding, speech recognition, text generation, machine translation, named entity recognition, Sentiment Analysis, text summarization, and question answering. In text understanding, machines are trained to extract entities, relationships, and sentiment from text, while speech recognition converts spoken language into written text, enabling machines to comprehend and respond to verbal commands.

In this section, we leverage NLP text preprocessing techniques to comprehensively analyse the articles presented to the user. Text Analytics, a subset of NLP studies, involves the automated extraction of valuable insights and patterns from unstructured text data. Its primary objective is to transform textual information into a structured format that can be subjected to various analyses. This field incorporates a range of techniques to decipher the linguistic and contextual aspects of textual data. As a first phase of the analyses, we identify whether the articles provided by Google News are really about that particular topic, therefore whether Google News "noticed" that the user was really interested in a topic. Second step is to analyse the content of the article, that is, analyzing the use of words and terms, how it is exposing the primary topic, or using tones against a particular theme, or if favorable. In the next sections, we illustrate the main components of Text Analytics. We base our study on a sample of 4 users, in particular from U_1 to U_4 (users with a particular political leaning). We analyse the first 10 articles presented on the For You page for each of them, so that we can analyse the different aspects of the articles presented to the user in the section, on 3 different days. The articles are stored in an XLSX [1] file format: we exploit the Python libraries to extract the links (as explained in the previous sections) of the articles, carry out a text preprocessing operation, and provide the article as input for the analyses. In addition, we take into account the position in which the first article relating to the user's primary topic is presented. The results will be presented in Chapter 7.

6.3.1 Sentiment Analysis

Sentiment Analysis is an NLP technique designed to determine the sentiment or emotional tone conveyed in a piece of text. The primary goal of Sentiment Analysis is to categorize the expressed opinions in the text as positive, negative, or neutral, providing valuable insights into the subjective view-

[1] Excel spreadsheet or worksheet.

points of individuals. This analytical process involves the use of algorithms and machine learning models to automatically analyse and interpret sentiments expressed in various forms of textual data, such as social media posts, customer reviews, news articles. In the context of social media, Sentiment Analysis can be employed to assess how users feel about a particular brand, product launch, or social issue. By evaluating the sentiments expressed in customer reviews, companies can gain actionable insights to improve their products, services, or overall customer experience. The process of Sentiment Analysis often involves pre-processing the text, tokenization (breaking the text into individual words or phrases), and then using machine learning algorithms to classify the sentiment of each text segment. Techniques range from rule-based systems to more sophisticated machine learning models that can discern nuanced sentiments. For our analysis, we exploit four different sentiment metrics: the Stanza library, Vader, MPQA and SentiWordNet.

Stanza is an NLP library for Python that provides a suite of pre-trained models and tools for a wide range of NLP tasks. The SentimentProcessor of the library adds a label for sentiment to each Sentence. The above methods identify multi-word tokens, which are then further extended into the syntactic words as the foundation for downstream processing. This is accomplished by the use of sequence-to-sequence (seq2seq) model to ensure frequently observed expansions in the training set, as they are always robustly expanded while maintaining the flexibility to model unseen words statistically. For each word in a sentence, Stanza assigns it as a part-of-speech (POS), and evaluates its universal morphological features (UFeats, e.g., singular/plural, 1st/2nd/3rd person, among others). To predict POS and UFeats, researchers adopted a bidirectional long short-term memory network (Bi-LSTM[2]) as the basic architecture. The pipeline consists of models ranging from tokenizing raw text to performing syntactic analysis on the entire sentence. The design is devised keeping the diversity of human languages in mind by data-driven models that learn the differences between languages. Besides, the components of Stanza are highly modular and reuses basic model architectures, when possible, for compactness. However, The existing models each support negative, neutral, and positive sentences, represented by 0, 1, 2 respectively.

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based Sentiment Analysis tool that is specifically designed to handle Sentiment Analysis for social media text. The VADER library uses a combination of sentiment lexicon (a list of words and their valence scores)

[2]Bidirectional processing that allows the model to capture both past and future context information simultaneously, which can be beneficial for tasks such as sequence labeling, sequence classification, and sequence-to-sequence learning.

and rule-based approach to analyse the sentiment of a piece of text. Unlike traditional Sentiment Analysis tools that use machine learning techniques, VADER doesn't require any training data to analyse sentiment. Instead, it uses a set of pre-defined rules and patterns to determine the sentiment of a piece of text. It is particularly useful for analyzing the sentiment of short and informal texts, such as tweets, online reviews, and chat messages. It takes into account both the polarity and intensity of the sentiment in the text, which makes it more accurate than traditional Sentiment Analysis tools in some cases.

MPQA (Multi-Perspective Question Answering) assigns each word in the lexicon a polarity score of either positive, negative, or neutral (i.e., a score of 1, -1, or 0, respectively) based on the word's emotional connotation. To exploit MPQA, we first load the MPQA lexicon (the file .tff[3] provided by the creators) and preprocess it, so that we could easily look up the polarity score of each word in the lexicon. Then, we define a function that takes a file path as input, reads in the text data from the file, and calculates the subjectivity score for each word in the text by looking up the polarity score of the word in the MPQA lexicon. The average subjectivity score that we calculate is simply the average of all the subjectivity scores of the words in the text. The maximum and minimum subjectivity scores represent the most and least subjective words in the text, respectively. Finally, the standard deviation of the subjectivity scores measures how much the subjectivity scores vary from the average subjectivity score, and thus gives us an indication of the overall degree of subjectivity in the text.

Sentiment Analysis using SentiWordNet involves using a lexical resource called SentiWordNet to perform Sentiment Analysis on text data. SentiWordNet is a publicly available lexical resource that assigns a sentiment score to each synset (set of synonyms) in WordNet, a large English lexical database. To perform Sentiment Analysis using SentiWordNet, the text data is first preprocessed to remove any noise and convert it into a format that can be analysed. Then, each word in the text data is assigned a synset based on its meaning. The sentiment score of each synset is then retrieved from SentiWordNet, and a sentiment score for the entire text is calculated by aggregating the scores of all the synsets in the text. The sentiment score can be used to determine the overall sentiment of the text, such as whether it is positive, negative, or neutral. This can be useful for a wide range of applications, including social media monitoring, market research, and customer feedback analysis. However, it is important to note that SentiWordNet is

[3]TrueType fonts are widely used for displaying text on screens and printing, and they are supported by various operating systems and applications.

based on WordNet, which is a English-centric lexical database, and may not be suitable for Sentiment Analysis of other languages.

Exploiting all of these sentiment metrics, we calculate the *average score* of article sentiments, calculated by the sum of the sentiment scores for each sentence, divided by the length of the sentiment scores. We calculate also the *maximum* and the *minimum* scores, and the *standard deviation*. This gives us an idea of how an article text in written, giving us a positive of a negative feedback. As a general overview, we have a pattern, for each user, of the situation about how Google News is providing articles with topics that are pro or either against a particular topic in which the user is interested. We calculate a total of 120 sentiment scores (10 for each day, and 3 days for each user, for a total of 4).

6.3.2 Readability Analysis

Readability Analysis is the evaluation of how easily a piece of text can be understood by readers. It involves assessing various linguistic and structural features of the text to determine its readability level. Factors such as sentence length, word complexity, and paragraph structure are considered in this analysis. Readability metrics, including formulas like the Flesch Reading Ease score, are often used to quantify these factors and provide a measure of how accessible the text is to readers. The goal is to ensure that written content is clear, comprehensible, and suitable for its target audience.

For our analysis, we exploit 8 different readability metrics, including Flesch Kincaid Grade Level and Gunning Fog, but for our final analysis we considered keeping only 5 for our final considerations. The 5 different Readabiity metrics are: Flesch Reading Ease grade, Dale Chall grade, ARI, Coleman Liau Index and Gunning Fog. The Flesch Reading Ease score is a measure of how easy a text is to read. It is calculated based on the average number of syllables per word and the average number of words per sentence. The higher the score, the easier the text is to read. The score ranges from 0 to 100, with higher scores indicating easier text. The Dale-Chall formula uses a combination of two factors to determine the readability score: the *average sentence length* and the *percentage of difficult words in a text*. Difficult words are words that are not among a list of 3,000 familiar words that are likely to be known by a fourth-grade student. The Automated Readability Index (ARI) is a readability formula that uses sentence length and word length to estimate the grade level required to read a particular text. The formula is: ARI = 4.71(characters/words) + 0.5(words/sentences) - 21.43. The Coleman-Liau Index is a readability test that uses a mathematical formula to calculate the approximate reading level needed to understand a given text.

The formula takes into account the average number of words per sentence and the average number of characters per word to determine the text's reading level. The Coleman-Liau Index produces a score that corresponds to a grade level, which indicates the minimum education level needed to understand the text. The score is based on a scale of 0 to 12, with 12 being the highest score and indicating that the text can be easily read and understood by a person with a 12th-grade education or higher. The score is calculated using the following formula: CLI = 0.0588 * L - 0.296 * S - 15.8, where L is the average number of letters per 100 words in the text, and S is the average number of sentences per 100 words in the text. Finally, the Gunning Fog index assigns a grade level to a text based on the number of complex words it contains. Complex words are those with three or more syllables, excluding proper nouns, compound words, and familiar jargon. The higher the number of complex words in a text, the more difficult it is to read, and the higher the grade level assigned. The Gunning Fog index formula is: Gunning fog index = 0.4 * ((total words / total sentences) + 100 * (complex words / total words)). The reason why we have not included all the metrics is because, obviously they differ in how they are built internally to be able to evaluate a text article, but they are all based on two main indices: the *score*, which is a numerical value that represents the metric score, calculated by taking into account the average number of words per sentence and the average number of syllables per word (the score is typically between 0 and 100, with higher scores indicating easier to read text), and the *grade level*, that estimates the reading level required to understand the text. It is expressed as a grade level (e.g. 4th grade, 8th grade, etc.) and is based on the metric score. A lower grade level indicates that the text is easier to read, while a higher grade level indicates that the text is more difficult to read. To exploit the readability metrics, we used the *Readability* Python's library, which is a tool that focuses on evaluating the readability of text by providing various readability scores. It includes all the metrics we named. Suppose we give an article as a parameter to calculate the readability score. Suppose we want to understand the Flesch Kincaid Grade Level of an article, and suppose the articles is about Political topics. Articles about Politics are much more difficult to understand than articles about Science, rather than Fashion or Sports. The metric in question could give us 90 as metric score (depending on the score. Usually, it could start from 0 to a maximum of 100), and evaluate the grade level as "Post-secondary reading, with highly specialized vocabulary and complex sentence structures. Academic and technical texts are typically at this level."

6.3.3 Part-Of-Speech (POS) Tagging

The Part-Of-Speech (POS) Tagging is a process in NLP that involves labeling each word in a text corpus with its corresponding part of speech, such as noun, verb, adjective, etc. NLTK provides several methods for POS tagging, including the default tagger, regular expression tagger, unigram tagger, and bigram tagger, among others. These taggers use various techniques, such as rule-based approaches, statistical models, and machine learning algorithms, to assign the appropriate POS tags to words in a text corpus. The accuracy of POS tagging depends on the quality of the training data and the effectiveness of the tagging algorithm used. POS tagging is a crucial step in many NLP applications, such as text classification, Sentiment Analysis, and information extraction, as it helps to identify the syntactic structure and meaning of a sentence.

The analysis starts tagging the sentences and identifying the number of adjectives in each sentence. For each article, we calculate the *total number of adjectives*, *total number of proper names*, and *total words*. POS tagging helps us in understanding the syntactic structure of sentences. By identifying the grammatical roles of words, it provides information about how words function within a sentence. Also, it helps disambiguate such words by considering their grammatical context, improving the accuracy of subsequent analyses.

6.3.4 Topic Modeling

Topic Modeling is technique to extract the hidden topics from large volumes of text. Topic model is a probabilistic model which contain information about the text. For example, a news paper corpus mcould contain topics like Economics, Sports, Politics, weather. Topic models are useful for purpose of document clustering, organizing large blocks of textual data, information retrieval from unstructured text and feature selection. Finding good topics depends on the quality of text processing, the choice of the topic modeling algorithm, the number of topics specified in the algorithm.

Gensim LDA's approach, used for the project, considers each document as a collection of topics and each topic as collection of keywords. Once we provide the algorithm with number of topics all it does is to rearrange the topic distribution within documents and key word distribution within the topics to obtain good composition of topic-keyword distribution. Topics are nothing but collection of prominent keywords or words with highest probability in topic, which helps to identify what the topics are about. The first step of the process is to create *bigrams* and *trigrams*. A bigram model is a language model that uses a history of one preceding word to predict the

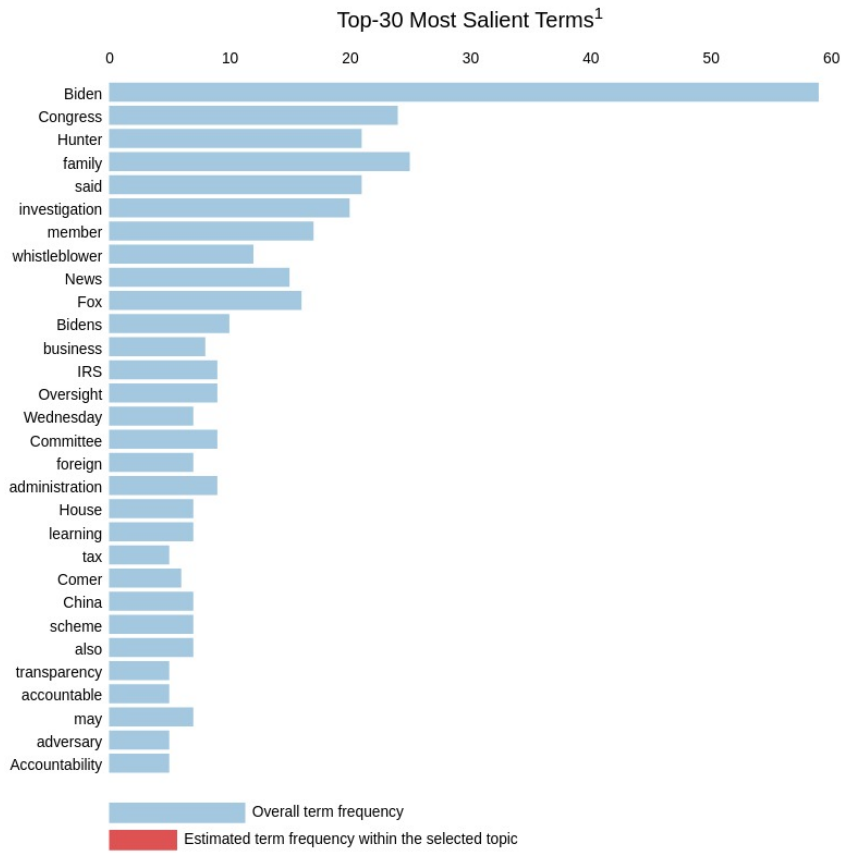


Figure 6.2: Topic Classification by the LDA Model.

next word. It is a type of n-gram model, where n is the number of words in the history. For example, a bigram model would predict the word "dog" given the preceding word "the" as "the dog". A trigram model, on the other hand, uses a history of two preceding words to predict the next word. For example, a trigram model would predict the word "jumps" given the preceding two words "the quick" as "the quick jumps". Both bigram and trigram models are used to improve the accuracy of tasks such as text classification. By incorporating more context into the model, they are able to better capture the meaning of the text and make more accurate predictions. Then, we create a Gensim[4] dictionary object id2word[5] from the list of preprocessed

[4]Python library for topic modeling, document indexing, and similarity retrieval with large corpora.

[5]Dictionary mapping word ids to words in the vocabulary. In other words, it's a mapping between the unique integer ids assigned to each word in the corpus and the actual words themselves.

and lemmatized texts (the dictionary assigns a unique id to each word in the corpus), the Gensim corpus object from the dictionary and the list of preprocessed and lemmatized texts. The corpus is a list of bags-of-words, where each bag-of-words is a list of tuples. Each tuple represents a term and its frequency in the corresponding document. Each tuple in the output represents a bigram that has been transformed into a two-element tuple. The first element of each tuple is the ID of the corresponding bigram in the dictionary (id2word) and the second element is the count of how many times that bigram appears in the input text. Each topic is combination of keywords and each keyword contributes a certain weightage to the topic. The top keywords and weights associated with keywords contributing to topic. Topics are words with highest probability in topic and the numbers are the probabilities of words appearing in topic distribution.

Let's take an example: suppose we identify the first sentence of an article as "*Republicans respond after IRS whistleblower says Hunter Biden investigation is being mishandled*". Corresponding bigrams are: [('Republicans', 'respond'), ('respond', 'IRS'), ('IRS', 'whistleblower'), ('whistleblower', 'says'), ('says', 'Hunter'), ('Hunter', 'Biden'), ('Biden', 'investigation'), ('investigation', 'mishandled'). The dictionary could be: [[(0, 1), (1, 1)], [(1, 1), (2, 1)], [(2, 1), (3, 1)], [(3, 1), (4, 1)], [(4, 1), (5, 1)], [(5, 1), (6, 1)], that is the first tuple (0, 1) represents the bigram ('Republicans', 'respond'), where Republicans has ID 0 in the dictionary, and respond has ID 1. The number 1 indicates that this bigram appears once in your text corpus. The top keywords and weights associated with keywords contributing to topic could be: *[(0, '0.004*"Biden" + 0.004*"mishandled" + 0.004*"Department" + 0.004*"either" + ' '0.004*"choice" + 0.004*"single" + 0.004*"every" + 0.004*"need" + ' '0.004*"potential" + 0.004*"behavior"'), (1,*. So for the first topic, the top 10 words and their weights are: "Biden": 0.004 "mishandled": 0.004 "Department": 0.004 "either": 0.004 "choice": 0.004 "single": 0.004 "every": 0.004 "need": 0.004 "potential": 0.004 "behavior": 0.004. This suggests that the documents in the corpus that are associated with this topic may contain discussions about Biden, the Department, mishandled situations, choices, and behavior, among other things.

Second step, we compute the *Perplexity* and *Coherence* scores. Coherence is a measure of how coherent the topics are. Higher coherence scores indicate more coherent topics. Perplexity is a measure of how well the LDA[6] model predicts the corpus. Lower perplexity scores indicate better predictions. Finally, we produce the LDA model: the visualization shows the topics

[6]Latent Dirichlet Allocation (LDA) is a generative probabilistic model in which each document in the corpus is assumed to be generated by a probabilistic process.

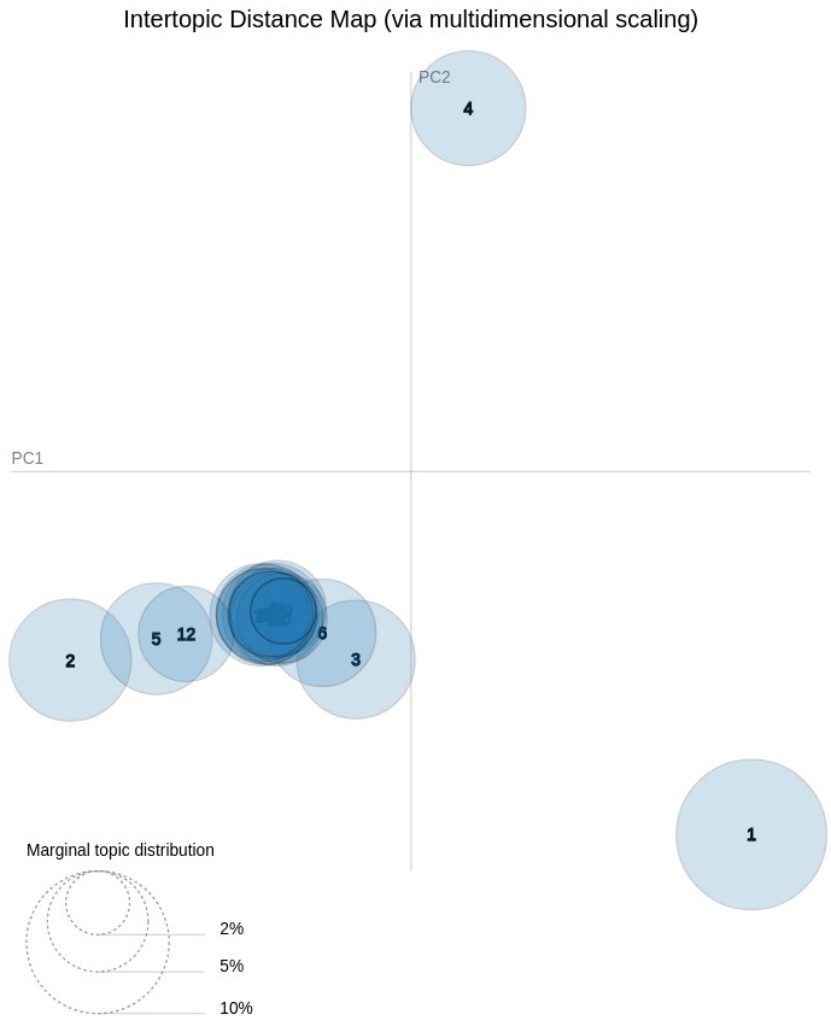


Figure 6.3: Intertopic Distance Map of Topics generated by LDA Model.

generated by the LDA model as circles, where the size of the circle represents the prevalence of the topic in the corpus. Each topic is represented by a list of words associated with that topic, and the strength of the association is represented by the distance between the words and the center of the circle. The visualization also shows the distribution of documents across topics, shown in Figure 6.2, where each document is represented by a horizontal bar chart. The length of the bar represents the prevalence of the document in the corpus, and the color of the bar represents the topics that the document is associated with. Each bubble on the left-hand side represents topic. The larger the bubble, the more prevalent or dominant the topic is. Good topic model will be fairly big topics scattered in different quadrants rather than

being clustered on one quadrant. The model with too many topics will have many overlaps, small sized bubbles clustered in one region of chart. Then, the Intertopic Distance Map plot visualizes the relationships between topics based on their similarity or dissimilarity, as shown in Figure 6.3. It is often represented as a heatmap or a scatterplot, where each point corresponds to a topic, and the distance between points reflects the degree of dissimilarity between topics. The intertopic distance map helps identify clusters of related topics and reveals the overall structure of the topic space. Topics that are closer together in the map are more similar in terms of their word distributions, while topics that are farther apart are more distinct from each other. In the intertopic distance map produced by Gensim's LDA topic modeling, the terms "PC1" and "PC2" refer to the first and second principal components, respectively. Principal component analysis (PCA) is a dimensionality reduction technique commonly used to visualize high-dimensional data in a lower-dimensional space while preserving the most important information. When applied to topic modeling, PCA is used to reduce the high-dimensional space of topic distributions (each topic represented by a distribution of word probabilities) into a two-dimensional space that can be easily visualized. In this two-dimensional space, each topic is represented by a point, and the distance between points reflects the similarity or dissimilarity between topics. The circles with the highest numbers inside (usually represented by larger circles or darker shades of blue) typically represent topics that are more central or prominent in the topic space; topics represented by larger or darker circles may signify overarching or dominant topics in the corpus, which are important for understanding the main themes or subjects covered by the documents. These topics may capture broad concepts or widely discussed issues that are central to the content being analysed.

6.3.5 Dependency Tree Height

Calculating the Dependency Tree Height of each sentence involves analyzing the syntactic structure of each sentence to determine how the words are related to each other. Dependency parsing is the process of analyzing the grammatical structure of a sentence and determining the relationships between words. Dependency tree length refers to the number of edges in the dependency tree that represents the sentence. The length of the dependency tree is an indication of the complexity of the sentence.

By finding the *average*, *maximum*, and *minimum* dependency tree length of a set of sentences, we can gain insights into the complexity of the text. For example, if the average dependency tree length is high, it may suggest that the text is more complex and harder to comprehend. Overall, tokenizing a

```
come
|  +--outcries
|  +--|  +--The
|  +--|  +--congressional
|  +--as
|  +--|  +--whistleblower
|  +--|  +--|  +--a
|  +--|  +--|  +--within
|  +--|  +--|  +--|  +--Service
|  +--|  +--|  +--|  +--|  +--the
|  +--|  +--|  +--|  +--|  +--Revenue
|  +--|  +--|  +--|  +--|  +--|  +--Internal
|  +--alleges
|  +--|  +--mishandled
|  +--|  +--|  +--investigation
|  +--|  +--|  +--|  +--an
|  +--|  +--|  +--|  +--into
|  +--|  +--|  +--|  +--|  +--Biden
|  +--|  +--|  +--|  +--|  +--|  +--Hunter
|  +--|  +--|  +--is
|  +--|  +--|  +--being
|  +--|  +--|  +--by
|  +--|  +--|  +--|  +--administration
|  +--|  +--|  +--|  +--|  +--the
|  +--|  +--|  +--|  +--|  +--Biden
|  +--.
```

Figure 6.4: Example of Dependency Tree.

text into sentences and analyzing the dependency tree length of each sentence is useful because it can help us understand the structure and complexity of a text, which can inform various NLP tasks and improve their accuracy. The depths dictionary contains the depth of each node in the dependency tree. The keys of the dictionary are the text of the nodes, and the values are the depths. We create a function that recursively traverses the syntactic dependency tree of each sentence in the article and stores the *depth* of each *node* in the depths dictionary. The depth of a node is defined as the *number of edges on the path* from the node to the *root* of the dependency tree. The function is called on the root of each sentence (*sent.root*) with an initial depth of 0. The depths dictionary maps each token (i.e., word or punctuation symbol) in the article to its depth in the dependency tree. The keys of the dictionary are the orthographic forms (i.e., the string representations) of the tokens, and the values are the depths.

For each sentence in an article, we calculate the *depth* (corresponding at the max tree depth), the words at the *max depth*, the *average tree length* (calculated by the sums of the length of each sentence, divided by the number of them), the *maximum* and the *minimum* tree length.

Chapter 7

Results

For our experiment, we include the readability score, number of topics, sentiment scores and lexical analysis of each recommended news articles. This provides us with a high-level, but above all complete, general vision of the use of certain words and sentences within each news article, thus providing us with the real unconscious meaning within the article. In this way, we are able to understand whether the article was able to hide biases (if present) that are unbalanced towards the theme of the topic, or the opposite effect, if against the theme presented in the article. We calculate each of these scores on the news articles recommended to an user, and, in order to ensure our observations are not biased by a particular day, we repeat the experiment for each user, and for 7 days.

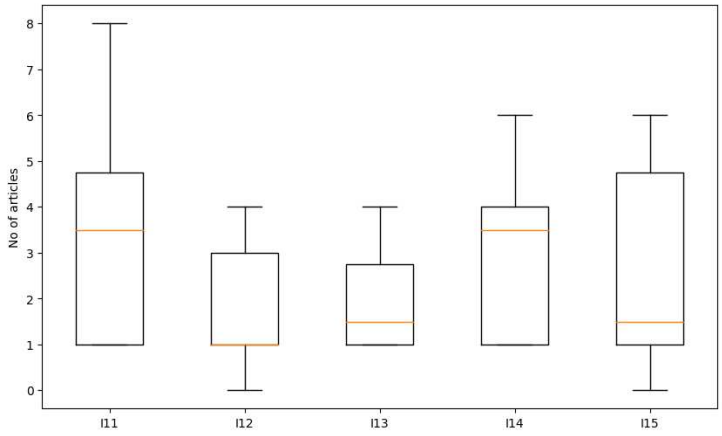
7.1 Readability Scores

In the previous subsection, we defined the readability scores as a score which gives a quantitative understanding of the difficulty of any text. For this study, we investigate whether the readability scores of the recommended news varies across users: we study whether an user is highly likely to be recommended more difficult news articles on the basis of his macro and micro news topics. We are able to see that the various readability analysis metrics, although differing based on the different levels of calculated scores, are not that far apart. Apart from the final readability calculation attributed to the news article, unique in each metric, we notice that they all provide a score, a numerical value which represents the score of the article, and a grade level, which represents the minimum level to fully understand the topic addressed in the article. After testing the various metrics for each article, and for each user, we notice that the score doesn't differ too much between the various

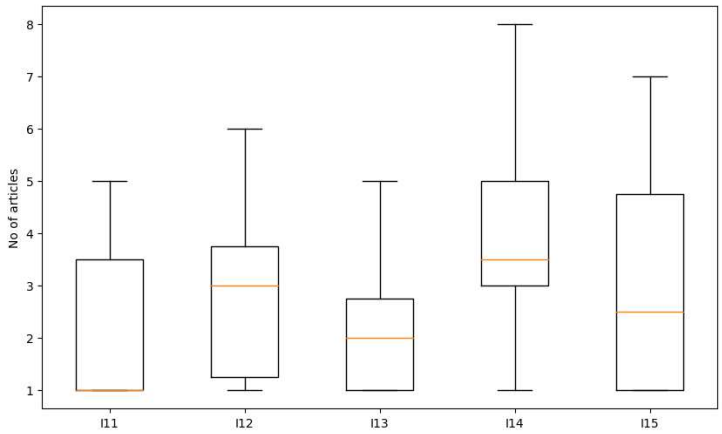
metrics (a score attributed to an article that talked about politics using Flesh Kincaid returned a college grade level, as well as Gunning Fog and many others). The score change little in the different metrics: there aren't metrics that, for a particular score, differ from the score of the others. In conclusion, for our experiment we decide not to include all the scores of each metric, because it is redundant. We have included the Flesch–Kincaid, Dale Chall, Coleman Liau Index and Automated Readability Index (ARI), respectively to calculate the readability score. Therefore, we conclude that the readability scores of an user has no pattern among users irrespective of their macro, micro news topic, location and the readability metric.

7.2 Topic Analysis

Understanding the number of topics in any text provides an understanding of the coverage of that news article. The topic analysis helps us understand how each news article provided to each user is composed, macroscopically speaking. In this experiment, we compare the number of topics in the recommended news and how it varies across users. We apply coherence based Latent Dirichlet allocation to identify the number of topics in a news article and we repeat this for all the users for a week. We show two representative examples where the box plot represents the number of topics of the news articles recommended to the users on two different days in Figure 7.1a and Figure 7.1b, respectively. Our observations indicate that there is low variance in the number of topics across users. We observe that the number of topics ranges between 1 to 4 on an average irrespective of the macro and micro news topic. We observe similar behavior irrespective of the day and location of the user. Although there are few news articles which has higher number of topics, such as, I_{11} and I_4 has one news article with 8 topics as shown in the figures, this is observed mostly as outliers.



(a) Day 1: 02 April 2023.



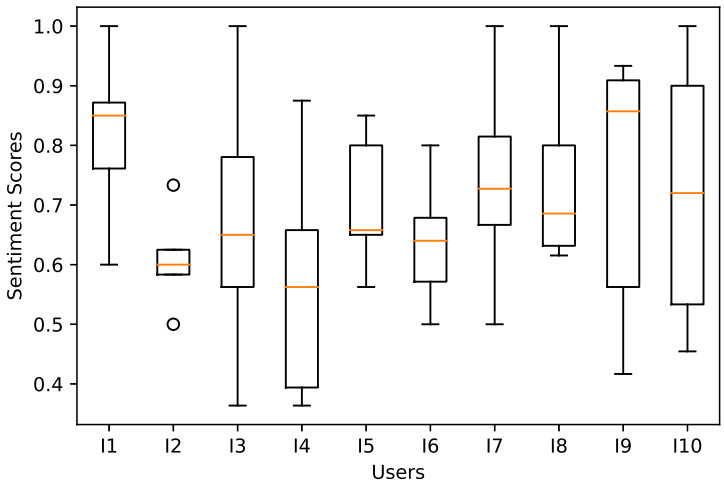
(b) Day 2: 03 April 2023.

Figure 7.1: Number of Topics in a News Article.

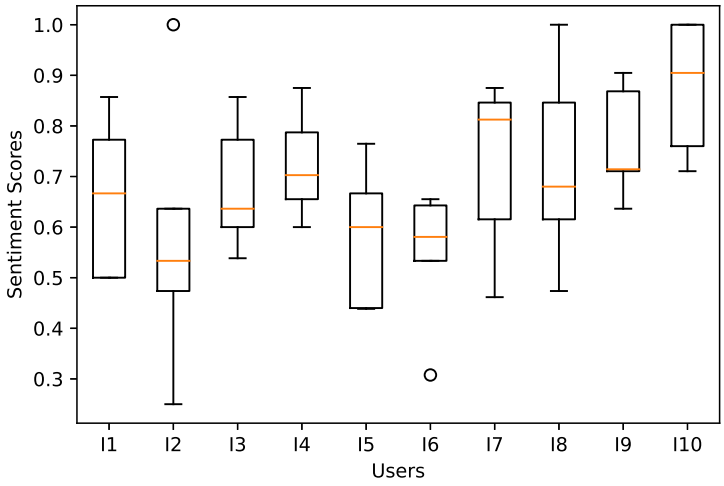
7.3 Sentiment Analysis

For Sentiment Analysis, understanding the polarity of the recommended news can provide an intuition about the variance in polarity in news article content across users. For this experiment, we calculate the sentiment of a news article on the basis of the sentiment score of the sentences present in the news article. We calculate the sentiment of a sentence by exploiting SentiWordNet and the sentiment score of a news article as the fraction of sentences in that news article which has sentiment. For example, an article with 0.7 sentiment score means that 70% of the sentences in that news article are non-neutral. Our observations indicate that a news article has generally more than number

of positive sentences followed by neutral sentences with very few negative sentences. Therefore, given k number of news articles recommended to an user, there is a high probability that the sentiment score of the majority of k news articles is higher than 0.5 due to the presence of positive sentences followed by few news articles which have score less than 0.5. We show our observations for 2 days in Figure 7.2a and 7.2b respectively where the box plot represents the polarity scores of the top 10 recommended news articles for a user on that day. We observed that users tend to receive more number of news articles which are positive than neutral on an average irrespective of their macro and micro news topic.



(a) Day 1: 07 November 2023.

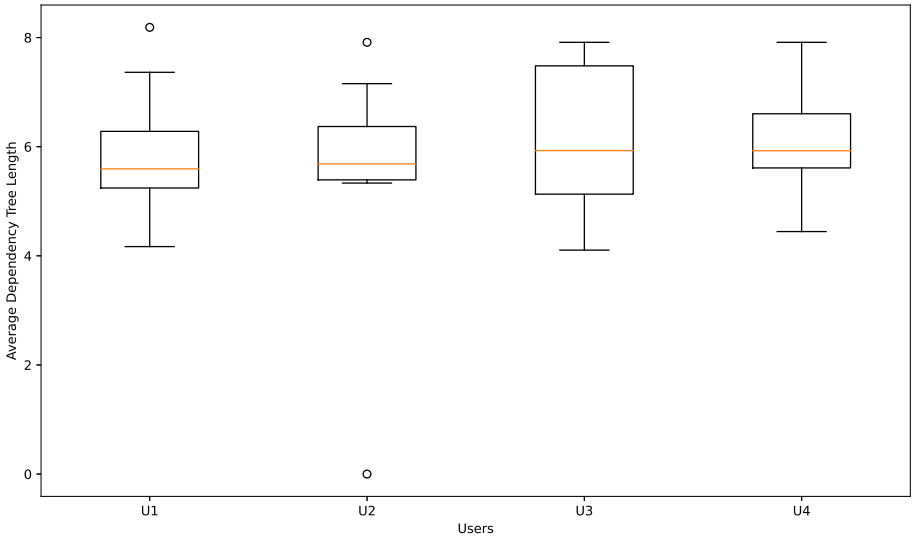


(b) Day 2: 08 November 2023.

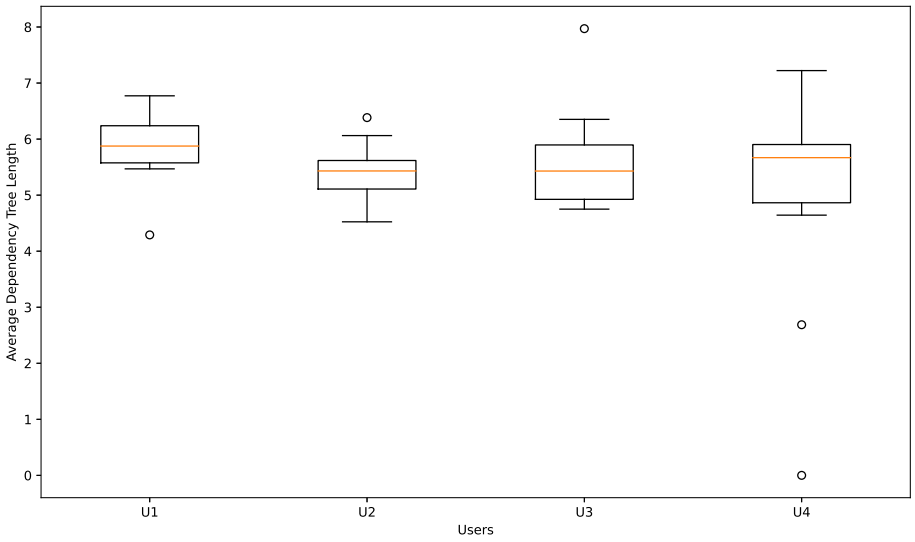
Figure 7.2: Sentiment Score for Users from India for 2 days.

7.4 Variance in Lexical Attributes

Finally, we study the variance in lexical attributes of a news article with respect to the macro and micro news topic. For this, we study the *dependency parse tree length*, *frequency of words*, *frequency of stop words* and *frequency of adjectives*, respectively for each recommended news article. In order to ensure our observations are not biased by a particular day, we repeated the same experiment for 7 days as we did for the previous experiments. Our observations indicate the average dependency parse tree length of the articles have very low variance, ranges from 4 to 6, irrespective of the macro and micro news topic, location and day of the news reporting. We show a couple of representative examples for different days for the same users in Figure 7.3a and Figure 7.3b. Our observations indicate that there is no relationship between frequency of words and frequency of adjectives with the macro and micro news topic, i.e., most of the news articles are similar in length except few outliers which does not give any significant relationship or correlation. Additionally, we don't find an user has higher likelihood to be recommended news articles with larger number of adjectives on the basis of the macro and micro news topic. We show the frequency of words and frequency of adjectives for the same day for same users in Figure 7.4b and Figure 7.4a. Therefore, our observations indicate that different forms of content analysis can not provide any critical insights to understand characterization of echo chambers for news media aggregators.

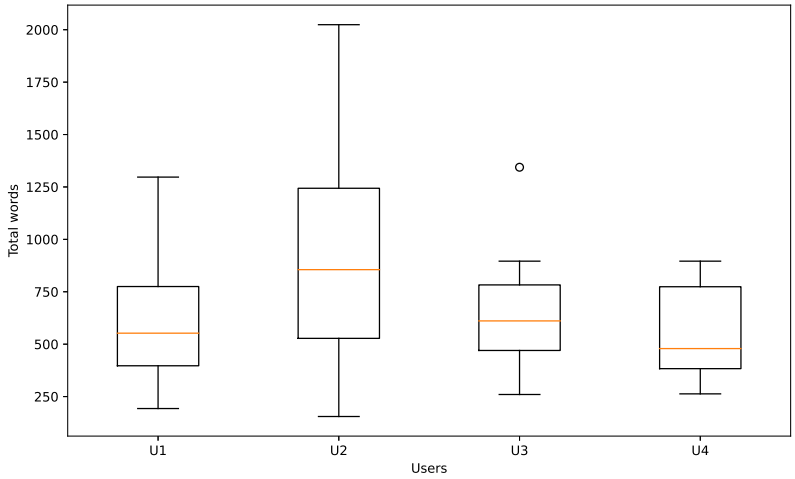


(a) Day 1: 02 April 2023.

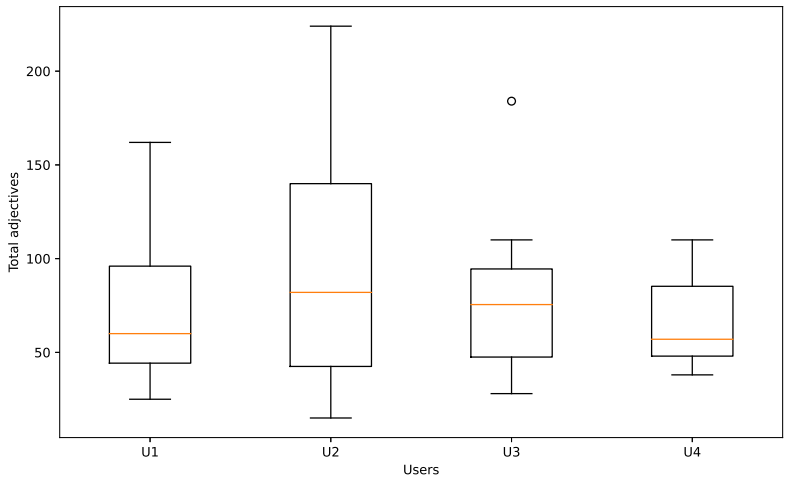


(b) Day 2: 03 April 2023.

Figure 7.3: Dependency Tree Height for Political Users from USA.



(a) Total Words of a Day of Analysis for Users from USA.



(b) Total Adjectives of a Day of Analysis for Users from USA.

Figure 7.4: Variance in Lexical Attributes.

7.5 Filter Bubbles

Filter Bubbles refer to the phenomenon wherein individuals are increasingly exposed to information and perspectives that align with their existing beliefs, preferences, and interests, while being shielded from contradictory or diverse viewpoints. In the context of news media aggregators, such as online platforms and social media networks, algorithms curate and prioritize content based on user-specific data, including past behaviors, interactions, and demographic information. As a result, users are presented with a personalized stream of news and information that reinforces their preconceptions, limits exposure to alternative viewpoints, and may contribute to the polarization of societal discourse.

Filter Bubbles are closely linked with the phenomenon of echo chambers due to their shared effects on information consumption and social interactions. An echo chamber refers to an environment, whether online or offline, in which individuals are predominantly exposed to ideas, opinions, and perspectives that reinforce their existing beliefs, values, and attitudes. The concept draws its metaphorical imagery from the way sound echoes within an enclosed space, amplifying and repeating itself. This phenomenom contribute to the creation and perpetuation of echo chambers by selectively exposing individuals to information that aligns with their preferences and viewpoints. When individuals are consistently presented with content that reaffirms their existing beliefs while filtering out dissenting or alternative perspectives, they are more likely to gravitate towards sources and communities that mirror their own ideologies. This process creates a reinforcing feedback loop wherein individuals become increasingly insulated from diverse opinions and dissenting viewpoints. The combination of Filter Bubbles and echo chambers can have significant implications for societal discourse and public opinion formation.

For this experiment, we take into account one day of analysis. If the results of the experiment had not returned what we expected, we would have repeated it another day, but the results satisfied us. For that particular day, and for each user from both the USA and India, we log into their user section, and take note of how many news about a particular topic Google News returned to that particular user. To do this, we build two matrices: the first m_1 x n_1 matrix, with m_1 rows as many as the topics of users from the USA, and n_1 columns as many as the users from USA plus a column dedicated to the Home section (the usefulness of the additional column will be explained in the next rows), and a second matrix m_2 x n_2, with m_2 rows as many as the topics of users from India, and n_2 columns as many as there are users from India plus a column dedicated to the Home section. The sum of each column is equal to 10: for each user, we take note of how many news articles up to a

maximum of 10 were presented by Google News to the users. In this way, we can define the numbers in the cell as the number of news articles presented to user i (with i ranging from U_i to U_m) belonging to topic j (with j ranging from T_1 to T_n). The first three topics of matrix 1 are, respectively, Republican Party, Democratic Party and Neutral Party. The first three topics of matrix 2 are, respectively, pro-Government Party, pro-Opposition Party and Neutral Party. The last column represents the Home column, i.e. the number of news articles belonging to a particular topic present on the Google News Homepage. This allows us to verify the importance of the topics assigned by Google News. It is important to underline this provision at least on the three political topics because they are the key point in the creation of the indices that allow us to identify the Filter Bubbles. The identification of Filter Bubbles within news media aggregators was important to understand how the algorithmic selection of news can influence users' opinions and perceptions, leading them to be exposed mainly to content that confirms and reinforces their pre-existing opinions, while ignoring or minimizing alternative points of view. The choice to focus on political topics such as the three topics considered is because the influence of Filter Bubbles in the political sphere can have significant consequences on society as a whole. Furthermore, political topics are often characterized by significant polarization, with clear divisions between various political factions. This polarization makes political topics particularly susceptible to the effect of Filter Bubbles, as people tend to seek confirmation of their political opinions and avoid information that might challenge them.

The experiment is based on these two matrices, from which we obtain two potential indices for the study on Filter Bubbles. The first is called *Average News Document Stance*: this index represents the average position of the news viewed by users based on the various topics considered. This index is calculated for each user and for each topic of interest. It's calculated for each user as a weighted average of the scores relating to the various topics. The first three rows of the matrix are extracted, which represent the topics of interest for the users (Republican Party, Democratic Party and Neutral Party for users from USA, pro-Government Party, pro-Opposition Party and Neutral for users from India), iterating through the columns of the matrix and for each user, we create a dictionary which contains the scores relating to the various topics. For each topic, the relative score for the user is calculated by dividing the number of news articles related of that topic viewed by the user by the number of total news article viewed by the user across all topics. This represents the fraction of news relating to that topic compared to the total news viewed by the user. For each user, it represents the distribution of the user's preferences with respect to the various topics, calculated as a

weighted average of the scores relating to the individual topics. This provides an indication of the user's average position with respect to the different topics considered. Therefore, we create three indices for each user, namely the Average News Document Stance (already presented in the previous section) regarding the Republican Party, the Democratic Party and Neutral Party for users from USA, and three indices regarding the pro-Government, pro-Opposition and Neutral party.

The second calculated index is called *Entropy User Score*. The calculation of user entropy (entropy user scores) evaluate the diversity of that particular user's preferences with respect to the various topics considered in the context to of the news aggregator. Entropy measures the uncertainty or variability of user preferences across different topics. To calculate this index, we scroll through the columns of the matrix (representing the users), and for each user we calculate the fractions relating to the number of news articles associated with each topic compared to the total news viewed by the user, and the entropy variations are calculated for each topic, using the Shannon Entropy formula:

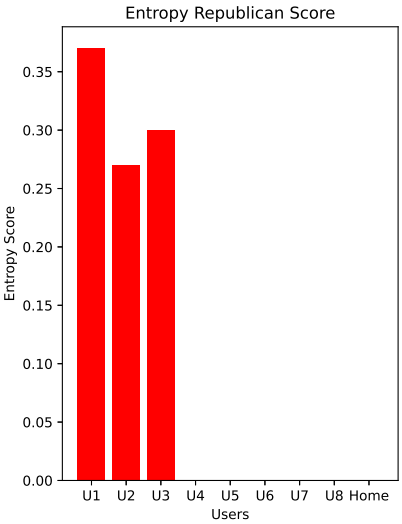
$$H(X) := \sum_{i=0}^{n} P(x_i) * log_2(P(x_i)) \tag{7.1}$$

where $P(x_i)$ represents the probability that the user sees news related to a given topic. If a user does not view any news on a particular topic, the relative entropy change is considered as 0. Finally, the entropy changes relating to all topics are added to obtain the total entropy of the user. User entropy provides a measure of the diversity or variability of their preferences with respect to the topics considered. Higher entropy indicates greater diversity in user preferences, while lower entropy indicates greater homogeneity or consistency in preferences. This measure can be useful for assessing users' tendency to be exposed to a broader range of topics or to focus on a limited number of specific topics. Therefore, these scores are two metrics used to evaluate different aspects of the diversity of users' preferences and the variety of news to which they are exposed. Together, they can provide useful information to identify the presence of Filter Bubbles in users of a news aggregator. The first index provides an overview of the average user preferences with respect to the various topics or themes. High uniformity in preference scores suggests that users are primarily exposed to content that confirms their pre-existing opinions. For example, if the majority of users have high scores for a particular political topic, such as Republican Party for USA users, it may indicate that users are exposed primarily to news that confirms their political leanings, while ignoring other points of view. The second index evaluates the

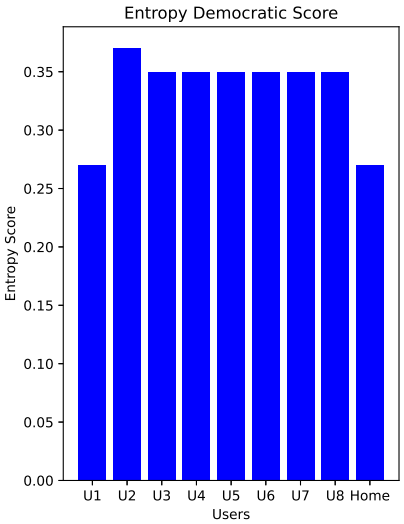
diversity or variability of user preferences with respect to the various topics. Lower entropy indicates greater homogeneity of preferences, while higher entropy indicates greater diversity. If most users have low entropy, it may indicate that they are strongly biased towards a certain set of topics. For example, if most users have low entropy regarding their political preferences, it might suggest that they are exposed primarily to news about a single political viewpoint and tend to avoid alternative topics or viewpoints.

Both U_1 and U_2 have higher scores for the Republican Party than the Democratic one and the Neutral. U_3 is inclined towards the Democratic Party and tends to view mainly Democratic news. Furthermore, it seems that news about the Democratic Party is not only present in users who view favorable news about the Democratic Party, but in all users who follow political news, plus the Homepage. The news aggregator may seek to provide a broader range of political news that reflects diverse perspectives, including Democratic Party views, regardless of users' political affiliation. Furthermore, given the diversity of news covered by the Democratic Party compared to the Republican one, Google News could try to balance users' exposure to political news from different sources and political perspectives, to avoid creating Filter Bubbles and guarantee informative coverage balanced.

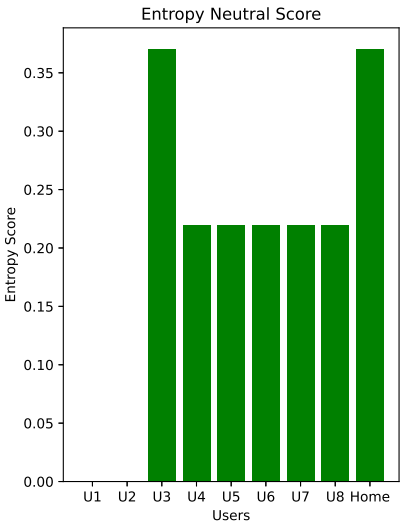
From the data provided on entropy scores for users in Figure 7.5 we can draw some observations. Users show a diversity of preferences regarding news related to the Republican Party (Figure 7.5a). In particular, some users have higher entropy, as evidenced by the values higher than 0.37. This indicates that these users have a wider variety of preferences when it comes to republican news, being able to view and engage with a more diverse range of viewpoints within this topic. In contrast, some users have an entropy of 0, which suggests that they may be more focused on a single aspect or perspective regarding the Republican Party, showing a more homogeneous preference in this area. Most users have a fairly uniform entropy for Democratic Party (Figure 7.5b), with values fluctuating around 0.35 and some values as low as 0.27. This suggests that users tend to have a variety of preferences when it comes to democratic news, albeit with a slight difference compared to republican preferences. This uniform entropy could indicate that users are exposed to different perspectives within the democratic sphere, without a clear bias towards a particular point of view. Finally, when it comes to neutral news (Figure 7.5c), users have different degrees of diversity in their preferences, with some being more likely to explore a wider range of neutral topics, while others may be more focused on specific areas of neutral interest.



(a) Republican Entropy Score.



(b) Democratic Entropy Score.



(c) Neutral Entropy Score.

Figure 7.5: Entropy Score for Users from USA.

Conclusions

With the shift in news consumption from printed newspapers to online forms, an user has continuous access to news throughout the day and her/his attention shifts on the basis of her/his choice. Therefore, to retain user attention, news media aggregators provide personalized reading recommendations which might often lead to formation of echo chambers, i.e., an user has access to news that aligns with her/his viewpoint. This has several side effects on society, like segragation among users, polarization, etc. Although several research works have been proposed to identify echo chambers in social media platforms, to the best of our knowledge, there is no existing literature which has studies echo chambers on news media aggregators. In this paper, we study the possibility of echo chambers in Google News Recommender. In order to understand the impact of news consumption topical choices on the news recommendation and how it varies across users, different topics and different location, we propose several metrics which can measure *Homophily in News Consumption and News Recommendation* and *User Similarity Analysis*, respectively. To *Homophily in News Consumption and News Recommendation*, we propose three different scores to exhaustively and effectively capture how news recommendation is very specifically aligned with the particular news topical choice. Additionally, we propose three different indices to capture user-user relationship on the basis of news recommendation and furthermore, how user specific news recommendation shield information across users. Furthermore, we also perform empirical analysis to understand the journey of an user to be caught in echo chambers and her/his susceptibility to fake news. We also observe how existing different measures of content analysis are not adept for echo chamber detection in Google News Recommender. As a future direction, we intend to extend this work to other/his news media aggregators and include users from other/his countries. Although characterization and visualization of echo chambers is of high significance, it highlights a major flaw in news recommendation, i.e., requirement of development fair news recommender approaches for Google News Recommender such that formation of echo chambers is prevented.

Bibliography

[Aiello et al., 2012] Aiello, L. M., Barrat, A., Schifanella, R., Cattuto, C., Markines, B., and Menczer, F. (2012). Friendship prediction and homophily in social media. *ACM Transactions on the Web (TWEB)*, 6(2):1–33.

[Alatawi et al., 2023] Alatawi, F., Sheth, P., and Liu, H. (2023). Quantifying the echo chamber effect: An embedding distance-based approach. *arXiv preprint arXiv:2307.04668*.

[Angela M. Lee, 2015] Angela M. Lee, H. I. C. (2015). The rise of online news aggregators: Consumption and competition. *Hsiang Iris Chyi*.

[Aral and Zhao, 2019] Aral, S. and Zhao, M. (2019). Social media sharing and online news consumption. *Available at SSRN 3328864*.

[Baumann et al., 2020] Baumann, F., Lorenz-Spreen, P., Sokolov, I. M., and Starnini, M. (2020). Modeling echo chambers and polarization dynamics in social networks. *Physical Review Letters*, 124(4):048301.

[Cinelli et al., 2020] Cinelli, M., Morales, G. D. F., Galeazzi, A., Quattrociocchi, W., and Starnini, M. (2020). Echo chambers on social media: A comparative analysis. *arXiv preprint arXiv:2004.09603*.

[Cossard et al., 2020] Cossard, A., Morales, G. D. F., Kalimeri, K., Mejova, Y., Paolotti, D., and Starnini, M. (2020). Falling into the echo chamber: The italian vaccination debate on twitter. In *Proceedings of the International AAAI conference on web and social media*, volume 14, pages 130–140.

[Daejin Choi, 2020] Daejin Choi, Selin Chun, H. O. J. H. T. K. (2020). Rumor propagation is amplified by echo chambers in social media. *Scientific Reports volume 10, Article number: 310 (2020)*.

[Del Vicario et al., 2016] Del Vicario, M., Vivaldo, G., Bessi, A., Zollo, F., Scala, A., Caldarelli, G., and Quattrociocchi, W. (2016). Echo chambers: Emotional contagion and group polarization on facebook. *Scientific reports*, 6(1):37825.

[Diaz-Diaz et al., 2022] Diaz-Diaz, F., San Miguel, M., and Meloni, S. (2022). Echo chambers and information transmission biases in homophilic and heterophilic networks. *Scientific Reports*, 12(1):9350.

[Duseja and Jhamtani, 2019] Duseja, N. and Jhamtani, H. (2019). A sociolinguistic study of online echo chambers on twitter. In *Proceedings of the third workshop on natural language processing and computational social science*, pages 78–83.

[E. Gilbert and Karahalios, 2009] E. Gilbert, T. B. and Karahalios, K. (2009). Blogs are echo chambers: Blogs are echo chambers. *2009 42nd Hawaii International Conference on System Sciences, Waikoloa, HI, USA, 2009, pp. 1-10, doi: 10.1109/HICSS.2009.91.*

[Edwards, 2013] Edwards, A. (2013). The inclusion and exclusion of dissenting voices in an online forum about climate change. *eration*, 2(1):127.

[Flaxman et al., 2016] Flaxman, S., Goel, S., and Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly*, 80(S1):298–320.

[Freeman, 1986] Freeman, J. (1986). The political culture of the democratic and republican parties. *https://doi.org/10.2307/2151619*.

[Garimella et al., 2018] Garimella, K., De Francisci Morales, G., Gionis, A., and Mathioudakis, M. (2018). Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the 2018 world wide web conference*, pages 913–922.

[Garrett, 2009] Garrett, R. K. (2009). Echo chambers online?: Politically motivated selective exposure among internet news users. *Journal of computer-mediated communication*, 14(2):265–285.

[Ge et al., 2020] Ge, Y., Zhao, S., Zhou, H., Pei, C., Sun, F., Ou, W., and Zhang, Y. (2020). Understanding echo chambers in e-commerce recommender systems. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 2261–2270.

[Gienapp, 1856] Gienapp, W. E. (1856). The origins of the republican party, 1852-1856.

[Gilbert et al., 2009] Gilbert, E., Bergstrom, T., and Karahalios, K. (2009). Blogs are echo chambers: Blogs are echo chambers. In *2009 42nd Hawaii international conference on system sciences*, pages 1–10. IEEE.

[Hada et al., 2023] Hada, R., Ebrahimi Fard, A., Shugars, S., Bianchi, F., Rossini, P., Hovy, D., Tromble, R., and Tintarev, N. (2023). Beyond digital" echo chambers": The role of viewpoint diversity in political discussion. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 33–41.

[Interian et al., 2023] Interian, R., G. Marzo, R., Mendoza, I., and Ribeiro, C. C. (2023). Network polarization, filter bubbles, and echo chambers: an annotated review of measures and reduction methods. *International Transactions in Operational Research*, 30(6):3122–3158.

[Jiang et al., 2019] Jiang, R., Chiappa, S., Lattimore, T., György, A., and Kohli, P. (2019). Degenerate feedback loops in recommender systems. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 383–390.

[Kim, 2008] Kim, H. N. (2008). The phenomenon of blogs and theoretical model of blog use in educational contexts. *Computers Education Volume 51, Issue 3, November 2008, Pages 1342-1352.*

[Kossinets and Watts, 2009] Kossinets, G. and Watts, D. J. (2009). Origins of homophily in an evolving social network. *American journal of sociology*, 115(2):405–450.

[Levy and Razin, 2019] Levy, G. and Razin, R. (2019). Echo chambers and their effects on economic and political outcomes. *Annual Review of Economics*, 11:303–328.

[Ludovic Terren, 2021] Ludovic Terren, R. B.-B. (2021). Echo chambers on social media: A systematic review of the literature. *Creative Commons Attribution-NonCommercial 4.0 International License.*

[Morini et al., 2021] Morini, V., Pollacci, L., and Rossetti, G. (2021). Toward a standard approach for echo chamber detection: Reddit case study. *Applied Sciences*, 11(12):5390.

[Morio et al., 2020] Morio, G., Morishita, T., Ozaki, H., and Miyoshi, T. (2020). Hitachi at semeval-2020 task 11: An empirical study of pre-trained transformer family for propaganda detection. In *Proceedings of the fourteenth workshop on semantic evaluation*, pages 1739–1748.

[Nathan Honeycutt, 2023] Nathan Honeycutt, L. J. (2023). Political bias in the social sciences: A critical, theoretical, and empirical review.

[Pye, 2015] Pye, L. W. (2015). Political culture and political development. *Princeton University Press*.

[Saribay, 1960] Saribay, A. Y. (1960). The democratic party, 1946-1960. *Political Parties and Democracy in Turkey*.

[Sasahara et al., 2021] Sasahara, K., Chen, W., Peng, H., Ciampaglia, G. L., Flammini, A., and Menczer, F. (2021). Social influence and unfollowing accelerate the emergence of echo chambers. *Journal of Computational Social Science*, 4(1):381–402.

[Susan Athey, 2021] Susan Athey, M. M. . J. P. (2021). The impact of aggregators on internet news consumption. *WORKING PAPER 28746*.

[Tassabehji, 2003] Tassabehji, R. (2003). Applying e-commerce in business. *2003 - Sage Publications Ltd.*

[Thompson and Santos, 2023] Thompson, J. E. and Santos, E. (2023). Echo chambers as gravity wells. In *2023 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 848–857. IEEE.

[Vrijenhoek et al., 2021] Vrijenhoek, S., Kaya, M., Metoui, N., Möller, J., Odijk, D., and Helberger, N. (2021). Recommenders with a mission: assessing diversity in news recommendations. In *Proceedings of the 2021 conference on human information interaction and retrieval*, pages 173–183.

[Wang et al., 2020] Wang, X., Sirianni, A. D., Tang, S., Zheng, Z., and Fu, F. (2020). Public discourse and social network echo chambers driven by socio-cognitive biases. *Physical Review X*, 10(4):041042.

[Wolfowicz et al., 2023] Wolfowicz, M., Weisburd, D., and Hasisi, B. (2023). Examining the interactive effects of the filter bubble and the echo chamber on radicalization. *Journal of Experimental Criminology*, 19(1):119–141.

[Yingqiang Ge, 2020] Yingqiang Ge, Shuya Zhao, H. Z. C. P. F. S. W. O. (2020). Understanding echo chambers in e-commerce recommender systems. *SIGIR '20: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[Zannettou et al., 2018] Zannettou, S., Bradlyn, B., De Cristofaro, E., Kwak, H., Sirivianos, M., Stringini, G., and Blackburn, J. (2018). What is gab: A bastion of free speech or an alt-right echo chamber. In *Companion Proceedings of the The Web Conference 2018*, pages 1007–1014.