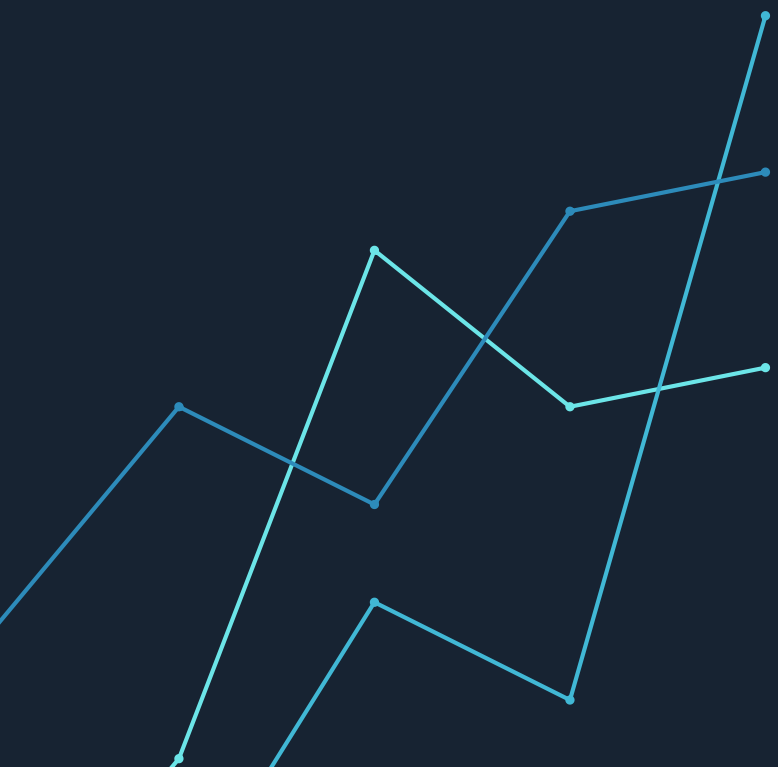
THE INTERNET

Analysis of the unobvious

By: Batsal Ghimire



Presenting

A Tool to uncover the secrets of the internet

Gihub Links

COMPLETE REPO

BatsalG/search-engine-app (github.com)

MAIN ANALYSIS

search-engine-app/SearchAnalysis/mainAnalysis at main · BatsalG/search-engine-app (github.com)

WEB APPLICATION

search-engine-app/SearchFullStack at main · BatsalG/search-engine-app (github.com)

BACKEND WITH AZURE

BatsalG/internet-bias-data: Azure SQL and Azure Functions to access data from Google, Bing and Twitter. (github.com)

To run the Github Repo, I have included the APIs and the Database Password in the Last Page of this paper. Rest of the instructions are in the Readme.md file in the repo.

EXECUTIVE SUMMARY

The Internet: Analysis of the Unobvious

By Batsal Ghimire

Search Engines and Social Media Platforms boast an aggregate base of more than 4 billion users and have established themselves as a critical source of information dissemination. More than 65% of the US population uses search engines as a gateway to access news, and two-thirds of them get it from social media alone (Shearer, 2021).

While operating at such scales, the editorial choices these platforms make by censoring certain topics have emergent effects in not only who they represent, but also what kind of inherent and unobvious traits are present in their top results. Rather than being a cynic and assuming malevolence in the intent of these platforms, this project is birthed from the acceptance that measuring "bias" is difficult without incorporating some significant unfounded assumptions. Instead, there is a need for a tool with which definite patterns can be deduced and presented for future researchers to build upon

AIMS

The project aims to develop an effective pipeline which allows the users to gather, access and analyze data about the sentiment and emotions of the articles and tweets, and create effective tools for visualizations and exploring patterns that reveal the trends in the sentimental, emotional, and political leaning for the platform ecosystem.

Instead of making claims about the real-world consequences of such findings, the strength of the project lies in providing data with minimal interpretations, so experts and researchers wanting to dig further can create their own hypothesis and test them using this platform.

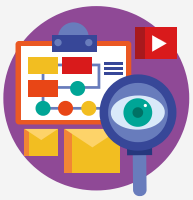
THE HARD PROBLEM IS...

Measuring Bias!

Rather than relying on expert assessments to categorize biases, this project is founded upon a comparative methodology across Google, Bing, and Twitter and uses Machine Learning to gather the parameters instead.

COLLECTION OF DATA

This project analyzes top results that are presented by search engines (Google and Bing) and social media sites (Twitter) for the 30 most controversial topics on the internet over the span of 4 months.



DATA PROCESSING

The project utilizes Natural Language Processing to extract the sentiment of the document, emotions (Joy, Sadness, Anger, Fear and Disgust), top keywords and entities along with its relevance.

Google Publishers

Among the top 10 publishers of Google, 7 are left-leaning and 3 are neutral, with no right-leaning publishers.

The New York Times, Reuters, and Al Jazeera are the top three publishers.

Bing Publishers

Among the top 10 publishers of Bing, 5 are left-leaning, 3 are neutral and 2 are right-leaning.

Fox News, The New York Times, and AP are the top three publishers.

Automating the process

To facilitate the research process, the paper also presents an application allowing for customization of keywords and automated data collection pipeline.

SENTIMENT

The average sentiment of Google, Bing, and Twitter are -0.22, -0.28, and -0.43.

Emotion

On average, sadness is the strongest emotion across the platforms, followed by joy, fear, and anger.

Adams, S. (2016, February 19). The Founder of DuckDuckGo Explains Why Challenging Google Isn't Insane. Retrieved from Forbes: https://www.forbes.com/sites/forbestreptalks/2016/02/19/the-founder-of-duckduckgo-explains-how-to-get-customers-before-you-have-a-product-and-why-challenging-google-isnt-insane/?sh=63e0db734e89

ABSTRACT

The ubiquitous nature of search engines and social media presents a challenge founded on the centralized nature of information sharing online, and the question on whether there are any individual and societal level consequences posed by such platforms is an ongoing subject of study. As such, building upon the literature on the homogenized nature of the internet and their role in fostering filter bubbles, this paper tackles the issue posed by the lack of topic-wise and temporal data to study traits of major digital platforms. By analyzing the sentiment, emotion, political leanings, and publishers of the results ranked the highest across Google, Bing and Twitter, and providing an intuitive and automated data collection and analytics pipeline, the project intends to renew the interest of academic and hobbyist researchers in understanding the differences in traits across these platforms, and hypothesize about the potential consequences of the patterns noticed.

Table of Contents

Search Engine and social media: Analysis of the unobvious

The most pivotal evolution in the way we consume and operationalize information has been brought upon by the rapid growth of the Internet, the process of navigating, which is made approachable by search engines and social media. And the impact of this remains omnipresent in our digital lives, with a total of 4.3 billion social media users and a comparable number of search engine users across the globe. Furthermore, the reliance on these channels to obtain information is at an all-time high, with nearly two-thirds of Americans getting their news from social media and 65% using search engines as a gateway to access news articles (Statista, 2022). On the surface, these staggering numbers might suggest a vast array of available platforms to satiate the ever-growing demand, however, a closer look reveals only a handful of platforms dominating their respective markets. For example, more than 90% of the search queries on the Internet are answered by Google alone (StatCounter, 2021).

Due to their size and scope, these platforms are commonly assumed to be totally automated and without human input, thereby assuring objective and credible results. However, this does not reflect reality since the algorithms and structure of these platforms are constantly modified, and undesirable topics like hate crimes, violence, etc., are actively avoided. Moreover, the topics presented are also fact-checked by third parties, which aids in the selection process. These editorial choices, however, might lead to certain content being favored over others, and this phenomenon is commonly referred to as "search engine bias." This bias isn't just limited to search engines, and it often creeps into social media as well, where the engagement generated by controversial topics incentivizes inauthentic private actors to generate and share polarizing material in attempts to influence the public opinions over controversial topics.

On the one hand, it is convenient to be cynical about the intent of these platforms without recognizing the fact that there are no perfect criteria to measure bias, nor is there such a thing as an objective opinion. Search Engines tweak their system to provide the greatest utility to the largest possible audience, and social media bridges the gap between information and entertainment, establishing both as an invaluable part of the digital ecosystem. On the other hand, the influences and complex interaction of the information across search engines and social media produces results that might have some unintended consequences. For example, an experiment conducted across a 20-year period on Facebook suggested the prevalence of emotional contagion in social media networks, which in some

cases changed users' emotions for days on end (Chen, Pacheco, Yang, & Menczer, 2021). With the understanding that objective measures of these traits are highly unreliable, this paper will present a unique comparative method of tackling the issue at hand. By examining the covert and overt influences of unobvious traits of content promoted across search engines (Google and Bing), and social media (Twitter), the paper will inspect the findings of the analysis for a set of 30 controversial internet topics over a span of 4 months. Furthermore, the paper will include a complete data-analysis platform that allows users to start their own study, either with existing data or new data of their choice.

In the first section, the paper will discuss the brief history of search engines and introduce the most popular ones in the market today. The second section will discuss the ranking algorithms used by Google and Bing, focusing primarily on PageRank. The third section will establish Twitter as an influential and uniquely positioned social media platform. The fourth section will then look at prior studies that discuss the social-psychological effects of these platforms. Finally, the paper will present the results of collected data and discuss the methodology and the applications of the developed pipeline. The aim of this paper is not to make a hypothesis about what potential impacts of the findings are, which actors are involved, whom it benefits, etc. It instead provides a strong dataset that can be used by targeted researchers to develop their own theories and give them a set of powerful APIs to create personalized datasets and visualizations.

PUTTING THINGS IN CONTEXT

BACKGROUND AND LITERATURE

History of the Internet and Search Engines
Page Rank Algorithm
Social Media & Twitter
Filter Bubbles
Individual and Societal Impacts

Internet and Search Engines: The survival of the fittest

How can information be transmitted during a Nuclear Attack? As an answer to this question lies one of the most important inventions in human history, the Internet. What began as a way for the US Defense Department to communicate with its contractors during the Cold War is now a household name, one without which we cannot imagine our lives (Featherly, 2016). Short for Advanced Research Projects Agency Network, ARPANET was the pioneer of the Transfer Control Protocol/Internetwork Protocol (TCP/IP), upon which the Internet was founded. TCP allowed computers to send data as small packets, and IP allowed these packets to efficiently route to their destination (Science Museum, 2018). This led to the rapid growth of Local Area Networks (LAN), and as the size of this network grew, it became hard to track and resolve IP addresses. In 1983, Paul Mockapetris and Jon Postel introduced the Domain Name System (DNS), which allowed the mapping of human-readable names into IP addresses, paving the path for the World Wide Web (WWW) (Pramatarov, 2018).

World Wide Web, invented by Sir Tim Berners-Lee, used HTTP (Hypertext Transfer Protocol) to transfer data formatted in HTML (Hypertext Markup Language) as web pages, which initially was only used by special-interest newsgroups (Web Foundation, n.d.). Soon, however, open-source directories were built, making it easier to navigate through the URLs, the most popular among which was the Open Directory Project (ODP) from Sun Microsystems (Carl Hendy, n.d.). Despite the popularity of ODP, it lacked the functionality to search using relevant keywords and descriptions, the void which was filled by Aliweb, the first web search engine (Nexor, 2014). Aliweb did not gain wider traction, and people still preferred to use directories to search for URLs. The commercial venture, WebCrawler, later became the first widely adopted search engine, which could index the entire webpage, enabling the users to search for any word or phrase within the content (Wikipedia, n.d.). Yahoo and Ask.com also leveraged keywords in combination with a manual ranking of the results and intuitive user interface to dominate the lucrative search engine market. The keyword-based ranking was not perfect, as it required the users to be precise with their queries. Over time, web admins started including irrelevant and repetitive keywords into the page's metadata, making the indexing less relevant to the intent of the query.

Robin Li, the founder of Baidu, tackled this problem by placing greater importance upon links from external websites to rank web pages instead of solely relying on keywords (Carl Hendy, n.d.). This was

also exploited, as people started creating web pages with many links pointing to other websites. Sergey Brin and Larry Page built upon this method by looking at the frequency of the backlinks and considering the importance and trust of the linking pages using their PageRank algorithm (Brin & Page, 2003). Google, which was built upon this, picked the low-hanging fruit when its competitors were struggling due to poor investments and slow innovation. The search results from Google were much more accurate compared to Yahoo and Ask.com, and within a few years, Google dominated the global search-engine market. Just three years into existence, Google began monetizing its platform through personalized advertisements, making it a revenue-collecting giant and leading Google to be one of the most profitable companies in the world (Rosenberg, 2020). In 2021, Google has more than 90% of the global search engine market share (StatCounter, 2021).

Alongside Google, Microsoft also aimed to take a pie of the search-engine market through its MSN Search, leveraging the fact that during the late 90s and early 2000s, more than 90% of Americans used Windows OS (Carl Hendy, n.d.). This did not go as planned, as MSN Search failed miserably. As a second attempt, in 2007, the service was rebranded as Live Search, which also failed to make any real dent in the market. It was again renamed Bing in 2009 when it partnered with Yahoo to power its platform moving forward (Gregersen, 2020). Despite this and several other improvements in their suite of services, Bing, in 2021, only accounts for around 3% of the global market (StatCounter, 2021).

Another recent contender is DuckDuckGo, launched in 2008 as a privacy-centric search engine. Initially, it held a very niche market among privacy-conscious users and failed to secure significant investments (Adams, 2016). However, after the leaks from Edward Snowden, exposing the data big-tech corporations were handing to the US Intelligence and highlighting that 98% of such data originated from Google, Microsoft, and Yahoo, the demand for a privacy-centric search engine rose rapidly, with even Safari and Firefox using DuckDuckGo as the default engine in 2014 (Carl Hendy, n.d.). Today, DuckDuckGo holds 0.66% of the global search engine market and 2.45% of the US market.

The Need for Change

The brief tenure of success demonstrated by platforms like Yahoo!, MSN Search, and WebCrawler can be traced back to its weakness in one of two key categories: resiliency to scalability and resilience to malicious users. Hierarchical navigation systems like Open Directory Project (ODP) maintained a

strong emphasis on having a high-quality set of sites listed in their directory and were constantly maintained and modified by editors (Couvering, 2008). With the advent of Web 2.0 and exponential growth of the number of sites, hierarchical schemes became difficult to navigate, and Graphs and Networks became a more suitable way to manage the growing arena.

A second series of contenders like WebCrawler, Yahoo, and Ask.com supported keyword search that indexed entire documents, greatly empowering the user experience and quality of results. The process also became largely automated and required less manpower to operate. However, malicious users started leveraging the weaknesses of these platforms by including irrelevant keywords in the documents, which forced the operators to manually edit the results, a process that soon became unsustainable. The need for innovation in the field was crystal clear, which was eventually provided by PageRank.

PageRank and its origins

If one were to pinpoint the most important innovation in search engines, PageRank would be the first in mind, which also simultaneously birthed the tech giant Google. Google, when introduced, had several enticing features that elevated it amongst its competitors, like increased storage efficiency by compressing documents and distributed crawling for Horizontal Scaling. However, Google tackled the heart of the problem that the search engines of that time were facing. Services like Yahoo! were heavily reliant on humans for indexing, making the results accurate but difficult to scale. Other engines that relied on keyword matching were being misled by advertisers and produced too many inaccurate results. PageRank tackled both these issues. The founding principle of PageRank was that "People were only willing to look at the first tens of results," which meant that the accuracy of the results was of paramount importance (Brin & Page, 2003). On top of that, the size of the index was growing rapidly, making keyword-based heuristics less precise. PageRank instead used link structure and anchor texts to filter out the results (Caren, 2021)[1].

Google's URLserver sends a list of URLs to the crawlers, which then fetches the data and compresses it before storing it in the database (Brin & Page, 2003). The indexer then decompresses the document, parses it, and stores it as word occurrences, referred to as hits. It also gets all the links from the web

[1] #heuristics Page 58

pages and stores information about its anchor text and other metadata, which helps identify where the links point to and from. It then uses PageRank to rank the web pages (Caren, 2021).

PageRank uses the number of citations for a webpage and the ranking of those web pages that cite it as a way to calculate the ranking of the page. If there are a lot of pages that are pointing towards it, then it has a higher PageRank (Gleich, 2015). Also, if there are a few pages pointing towards it, but those pages have a high PageRank, then it has a higher PageRank. This can be demonstrated mathematically using the following equation:

$$PR(A) = \frac{1-d}{N} + d(\frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \cdots)$$

Here, PR(n) is the PageRank of Page n, 'N' is the number of documents, L(n) represents the number of outbound links from the page. 'd' is the damping factor, which decreases the derived value, and after multiple tests, is usually considered to be 0.85 (Gleich, 2015).

A popular analogy to the PageRank is the Random Surfer, where a surfer at a certain page flips a coin to determine whether to click a link on that page or teleport to a random page irrespective of the current page's location (Wikipedia, n.d.). The web pages where the random surfer has a higher probability of lading have a higher PageRank. The PageRank of a page is also updated every time an index is rebuilt, so it preserves the relevance of a ranking. On top of PageRank, Google also considers factors like the font size, location of the keyword, capitalization, etc., to determine the relevance of the content (Brin & Page, 2003).

Although PageRank is still the foundation of Google Search, it upgraded its system to what is referred to as "Hummingbird" in 2013 (Shewan, 2014). The ranking factor now uses natural language processing to emphasize the context and meaning behind the search results rather than the keyword itself. It also allows Google to directly scroll to the most relevant part of a webpage, as it can detect the intent behind the query more effectively. This change has also enabled the developers to use more natural language terms instead of forcing keywords to the metadata.

Bing Search Algorithm

The algorithm behind Bing shares a lot of similarities with Google, however, the emphasis it places on the aspects of the web page varies. Bing places a much higher emphasis on the authority of the backlinks, i.e., it prioritizes results that have a highly ranked TLD (Top-Level Domain) (Smallwood, 2012). The indexing algorithm for Bing is also much slower compared to Google, which Bing says is to prevent spam as their spam detection is not as strong as Google's (Cornell University, 2018). So, once Bing establishes a list of high authority websites, it might not update it for weeks. This has two main consequences:

1. The results from Bing might be outdated compared to Google, which boasts a faster indexing algorithm and hence can confidently provide the most relevant and recent results.
2. Bing prioritizes well-established websites and '.edu' and '.gov' domains, which generally have a higher ranking. This can also mean less emphasis on newcomers and private blogs, along with social media sites or other public forums (Radd Interactive, n.d.).

There are similarities and differences between Google and Bing, both in terms of what they prioritize and the sophistication of their personalized results and natural language processing. However, some level of personalization and suggestion of search results are present in both, which might lead to some unwanted consequences.

Simultaneous brewing of social media

Most modern social media platforms that we currently recognize flourished in the mid-2000s, alongside the Web 2.0 landscape, which provided renewed interest in internet businesses after the Dot Com crash. However, the seeds of it were planted much before, when people reimagined community building, communication, and information transmission in the digital world. Usenet became one of the first and widely recognized discussion systems on the Internet, which built upon the technology used by email (K & Patil, 2013). The users could contribute and consume information that was categorized into groups, referred to as newsgroups. The next major step towards the advent of modern social media came through Internet Relay Chat (IRC), which built upon Usenet, and provided one-on-one private messaging. IRC was also used to disseminate information during the Soviet coup d'état media blackout and the Gulf War (K & Patil, 2013).

In 1994/95, private ISPs (Internet Service Providers) started commercializing the Internet in the US, and millions of people had access to these platforms, leading to its rapid growth. At the end of the century, however, social networking and blogging started garnering some interest, and there was a lagging base for UseNet and AOL Instant. Blogs allowed people to post about interesting and informational topics to web pages, which often had individual contributors, and the publishing process was made more appealing to non-technical audiences through web publishing tools. These articles were identified as posts and were showcased in reverse chronological order, making it an ideal way to communicate about a wide variety of recent topics (Boyd, 2015). Simultaneously, social networking sites like Classmates.com, Six Degrees, and Myspace were growing in their users, as they provided a fresh take on human interaction through their utilization of posts, messages, comments, image sharing, etc., all in a single platform. Myspace, in particular, grew rapidly, and today is widely regarded as the first modern social networking site, and since then, many others like Facebook, Instagram, Twitter, etc., have become an important part of our internet usage.

Twitter as microblogging and Social Networking Site

Among many platforms that flourished in the mid to late 2000s, Twitter is one of the few that integrates elements of blogging and social networking into a single platform. Created in 2006 by Jack Dorsey, Twitter started as a way to send short messages to large groups of people, often discussing real-time news and issues, and it grew from around 300,000 tweets per day in 2008 to more than 50 million by early 2010 (Meyer, 2020). Twitter has established itself as one of the top 3 social networking sites and a major way to stay up to date with the current global events (MacArthur, 2020). Unlike other platforms, Twitter has a strict limit of 280 characters, and contrary to Facebook, Instagram, and the likes, most accounts the users follow aren't people they know and interact with in real life. Instead, it acts like a blogging service, where users can share small parts of content, images, and links, which aggregated is commonly referred to as microblogging.

Now that we've discussed some of the most pivotal contexts behind the advent of the modern Internet, we will focus on what the consequences of this are in information consumption and dissemination.

The Internet and Filter Bubbles

Filter Bubble, a term popularized by internet activist Eli Pariser in 2010, refers to a state where an individual is isolated from views other than their own due to the personalization of the results from search engines and social media (Boutin, 2011). Many applications tailor their news feed and search results based on the user's past activity, search habits, followers, and the overall nature of their digital footprint, creating an echo chamber that reinforces their existing views and beliefs. Many legislatures are also concerned that Internet has become a breeding ground for misinformation, and some even fear that citizens making ill-informed decisions might harm democracy (DiFranzo & Gloria-Garcia, 2017). Features like auto-complete are also widely used, with 23% of the queries coming through it, where users are nudged towards the most popular and personalized search queries (Dean, 2020).

The evidence on the effects of search engines and social media on filter bubbles are mixed, however, with most stating that these platforms allow the users to turn off personalized feed if they want. Beyond that, many search engines like Google have decreased the level of personalized results on their pages. A study by NYU on Twitter's filter bubble stated that social media platforms, being more interactive, allows individuals to discuss political events with others, even with those they know little about (Barberá, 2014). Such discussions can lead to diversification of opinion but also can increase polarization. Another study that used social bots to introduce differing views to American Twitter users discovered that after the study, self-registered republicans demonstrated even more conservative views whereas the self-registered liberals showed no change whatsoever (Bail, et al., 2018). This challenges the fundamental premise of the argument, which states that the introduction of opposing opinions increases open-mindedness towards other ideologies.

And the trend toward lack of personalization in search engine results can also have negative consequences that are listed below:

1. There is a traffic drop of 95% from Page 1 to Page 2 of Google Search results, and the first result on the page sees 32.5% of traffic, a number which reduces to 3.5% for the 7th item (Digital Synopsis, n.d.). This highlights, in general, that if people don't find the answer they're looking for in the first tens of results, they consider the search to be a failure. So, with a lack of personalization, the search engines will cater to the likes of the majority, while minority likes and views (and webpages representing those) will gain limited exposure (Goldman, 2006).

2. Due to the low click-through rate in later results, websites and news publishers will be desperate to achieve a higher ranking and hence might promote materials with negative sentiments, generally known to increase polarization and social unrest. We already see this in the headlines, the majority of which cater to the negativity bias (affinity and increased arousal towards negative news) of people (Cherry, 2020). This can lead to anxiety, sad moods, and even mental health issues.

3. The existing power structure of web pages might get worse. PageRank considers webpages with high inbound links to be more authoritative, and if personalization is completely removed, large corporations with funding in marketing and Search-Engine Optimization (SEO) will dominate the search engine results pages (SERPs) (Cornell University, 2018)[2].

From reinforcing political beliefs to affecting the mental health of the users, Search Engine Results can have a wide array of effects, the magnitude, and direction of which is still not completely obvious. The unobvious and implicit traits of the search engines and the results that they present are interesting areas of study.

Impacts on individual and collective interactions

Ranking algorithms are beneficial tools to help clear out noise from search results, whether it is for a news topic or for a scientific article. And in this, popularity is a key metric that is used to determine what goes on the top for a given search query (Ciampaglia, Nematzadeh, Menczer, & Flammini, 2018). The 'wisdom of the crowd' method leverages the social influence theory, where an individual is more likely to engage in materials that are chosen by the people within their network. Assuming that these platforms successfully prevent bots and fake users from generating and influencing the showcased content, it still doesn't warrant an optimal query set. One major psychological theory that supports this is the cognitive load theory, where individuals use heuristics to make split-second decisions that aren't based on an accurate assessment of the material at hand but on less cognitively heavy metrics like popularity and emotional impulses (Islam, Laato, Talukder, & Sutinenb, 2020).

[2] #complexcausality: Page **3**

Another major aspect of this is what is referred to as 'echo chambers, where the attraction of social media users towards controversial topics like climate change, vaccination, etc., creates a homogeneous pattern in their online communities leading to radicalization and increased spread of misinformation (Garrett, 2009). As an illustration, there is evidence that suggests the recommendations provided by YouTube eventually lead to items with extremist viewpoints, regardless of what the initial starting point was (Ribeiro, Ottoni, West, Almeida, & Meira, 2020). The impacts of online behavior, unfortunately, aren't just limited to the digital world but have many real-world consequences. For example, the information consumed and interactions that people have on the Internet can affect the actions of the users and their emotions for many days (Kramer, Guillory, & Hancock, 2014)[3].

The next half of the paper will discuss a methodology with which we can compare some traits of the search results that aren't immediately noticed and discussed. It is to note, however, that these are measures of an entire ecosystem and not elements within it. Through this, we can assess patterns but not intentions and actors involved. For example, there is no reliable way to observe whether a particular pattern is a consequence of the decisions and algorithms used by the platform or if it is driven by the communities that use it.

For this, we will look into a comparative analysis of two search engines, Google and Bing, and a social media site, Twitter. There are several key metrics that are calculated and presented. The goal is to start a discussion about patterns that are observed rather than make predictions and calculations about their consequences. As an example: for a researcher wanting to study the impact frequency of publishers has on the sentiment or emotion of their articles, this tool will provide the metrics for them to do so. Starting a discussion on the impacts of sentiment, emotions, entities, keywords, etc., on the way humans make sense of information is a critical one and one that must be rooted in well-founded evidence and metrics. This is the exact goal of the tool that is presented.

[3] #conformity Page **59**

DEVELOPING A SOLUTION

DATA COLLECTION AND APPLICATION

Data Collection Pipeline
Application Development
Automating the Process
Feedback and Final Iteration

Data Collection

Topics that are the most aggravating provide more substantive evidence as to the differences in key characteristics of the platforms and often find themselves in media and political attention, making them key areas of study. To leverage this fact, the data will be collected for a set of 30 key topics of internet controversies. These topics have been chosen based on the "Wikipedia's Most Controversial Topics" list, which groups articles that are re-edited in a circular manner and are subject to most article sanctions (Wikipedia, n.d.). These topics are also further categorized into three issues, which are based on subjective evaluations. Arguments can be made about the cross-tabulation of these topics, and the tool provides an easy way for the researcher to do so. For example: Depression and Suicide can be considered social issues, however, it is categorized as scientific issues in this tabulation.

*TABLE 1: **THE TABULATION OF KEYWORDS ACROSS VARIOUS ISSUE TYPES.***

International Issues	Social Issues	Scientific Issues
China	Abortion	Artificial Intelligence
Israeli-Palestinian Conflict	Anti-LGBT	Suicide
Kashmir	Anti-Muslim Violence	Climate Change
Taliban	Extremism	Critical Race Theory
Nuclear Weapons	Feminism	Coronavirus
Russia-Ukraine (*Later Added in January 2022*)	Gay Marriage	Depression
	Gender Inequality	Marijuana
	Gun Control	Vaccines
	Hate Crimes	
	Immigration	
	Online Censorship	
	Police Brutality	
	Privacy Rights	

	Racism Religious Freedom Social Media	

Keyword Choice

Since it is a comparative analysis, the choice of keywords isn't of major concern for this use case. However, the platform allows easy incorporation of new keywords. For example: if someone wants to study the sentiment of results that come up when they search for Anti-Muslim vs. Anti-Hindu, they can easily do so. And, if they want to see the impact of different framings of keywords for the same topic (ex. Why climate change is fake? Vs. Why climate change is not fake?), it can be easily accomplished as well.

Note: In the analysis, there are 31 keywords. The 'Russia-Ukraine' keyword was added in early January and hence only consists of half the time duration as other keywords. Also, the data for Twitter only exist for a period of 2 ½ months.

The data is collected through a Python script that collects the top 10 results every 24 hours from November to December and every 12 hours from January to February. The repetition of the results will automatically be eliminated from the database.[4]

For each of the keywords, the Python Script for Bing and Google will extract the top 10 results from Bing and Google News, and for Twitter, it will use the Twitter API V1 to gather the most popular tweets for a given query. The new V2 API from Twitter doesn't yet support the search for popular results, so a valid API for V1 is needed, which requires elevated access from Twitter.

The following information will be stored for Google and Bing:

1. Article Title
2. Article URL

[4] #variables Page. 57

3. Publisher Name
4. Description of the article
5. Published Date

(The entire contents of the data are available from November to December. Post-December, fetching entire results is only used as a failsafe in case the API cannot directly read an article from the URL.)

For Twitter, the following information will be collected:

1. Tweet ID
2. Tweet Text
3. Creation Time of the tweet
4. Favorite Count
5. Retweet Count
6. URL for the tweet

Once these data points are stored in the database, we will be performing sentiment analysis on each article. We will be using IBM Watson's Natural Language Understanding Library, which has been tuned and is available through the API endpoint (IBM Cloud, 2021).

We will then extract the following attributes from the API:

1. **Document Sentiment**: A 0 to 1 value, where a negative value denotes a negative sentiment, and a positive value denotes a positive sentiment. Note: This sentiment is not towards a particular keyword and is a general sentiment for the entire document.

2. **Emotion**: There are five categories of emotion that we'll receive:

    a. Sadness

    b. Joy

    c. Anger

    d. Fear

e. Disgust

We will be referring to 'Sadness,' 'Anger,' 'Fear,' and 'Disgust' as negative emotions and 'Joy' as a positive emotions. The values for these range from 0 to 1, with a higher value denoting a more pronounced expression of that emotion.

There are further sentiment data that will be extracted by the Web Application but not included in the analysis. These can be accessed in the dashboard:

1. **Keyword Emotion**: 5 Emotions towards the search keyword
2. **Keyword Sentiment**: Sentiment towards the search keyword
3. **Top Keywords**: The 5 most relevant keywords and their relevance value and sentiment
4. **Entities**: Top 3 entities (name, organization, location, etc.) from the article and their relevance and sentiment.

During the analysis, we will also be looking at which publishers are the most represented in the search results, and then we will further look at their political leanings. This data is gathered from AllSides, which aggregates data through surveys and public perception responses, and categorizes the publishers as (Allsides, n.d.):

1. Left
2. Left-Center
3. Center
4. Right-Center
5. Right

All this is stored in a SQL Server Database, and the complete Database schema is shown below:

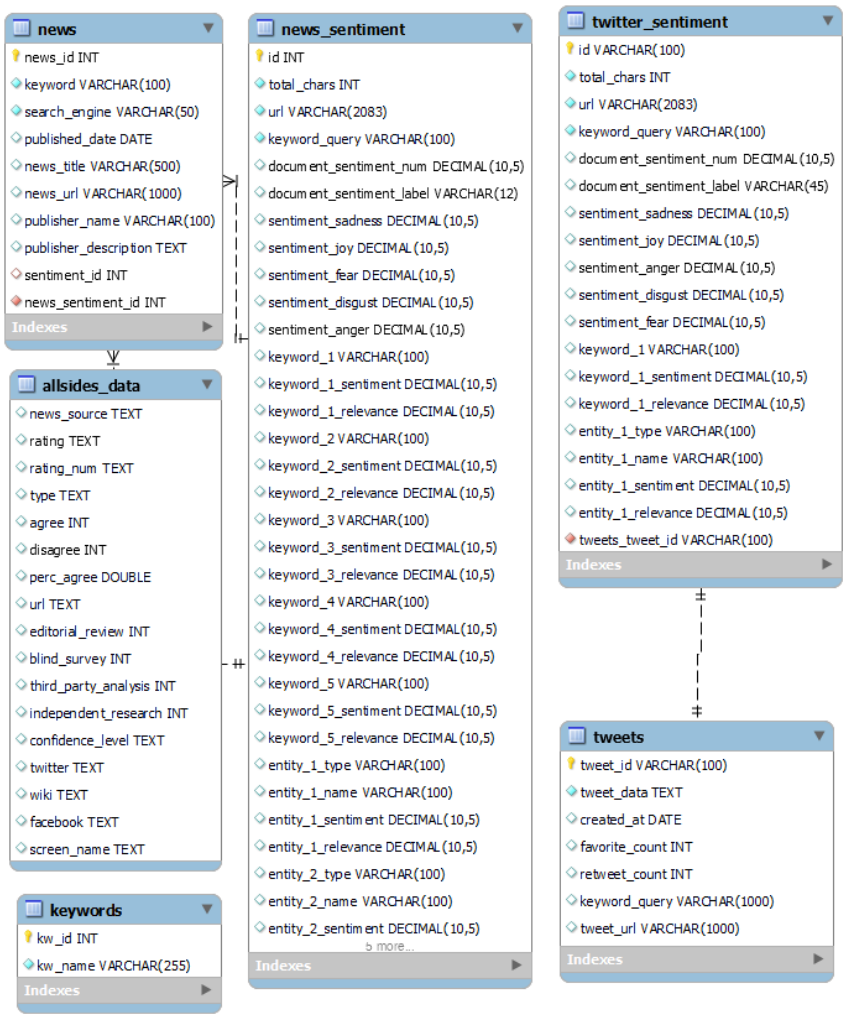


FIGURE 1: ***FIGURE: DATABASE SCHEMA FOR TWEETS AND NEWS AND ITS SENTIMENT.***

Main Functions

Gathering all these information has been abstracted away into a single function, each for Twitter and the Search Engines. They are:

```
fetch_from_engine(keywords_from = 'my_sql', number_of_results = 10)
fetch_from_twitter(keywords_from = 'my_sql', number_of_results = 10)
```

Both these functions take in two parameters:

1. **keyword_from:** Defaults to 'my_sql,' i.e., it goes through all the keywords that are included in the 'keywords' table in the SQL database.
   It can also take in strings or lists as an argument, which could be the keywords of interest.
   A string can be used for individual searches; a list can be used for both individual and multiple searches.
2. **number_of_results:** Total number of results to fetch from the engines or from Twitter. Defaults to 10. It can take in any integer value as a parameter.

There are multiple things that this function does under the surface:

1. **For Google and Bing:** It makes a call to the scraper to fetch all the necessary results and attributes and stores it in the database.
   For Twitter: It makes a call to the Twitter API and fetches the required number of tweets, and stores them in the database.
2. It then makes a call to the IBM API with the URL (For Google and Bing) and the tweet text (For Twitter) and fetches all the attributes, and stores them in the database.

(**Note**: In order to reduce the number of calls to the IBM Server, after fetching the search results, it checks if the information is already present (i.e., identifies duplicates) and does not insert the data into the sentiment tables.)

There are many other data cleaning and management that happens under the hood, but these are unimportant for the client to know. (The source code showcases all these details)

The tool also comes with a client-side application and a RESTful API with which the users can view and compare the search results across Google and Bing, along with a dashboard that visualizes the sentiment data.

Application Dashboard

The Application is developed with a Flask Backend and a React.js frontend. There are a few reasons behind this choice of the stack, which are mentioned below:

1. **Flask**
    a. Written in Python, Flask can help reuse the code in both the backend and the analysis, supporting the DRY (Don't Repeat Yourself) principle in software development.
    b. Flask provides a very minimal and straightforward way to create a RESTful API and creates great flexibility in handling the requests.
2. **React.js**
    a. Reusability of the components and logically separating them out into different files is the major reason for choosing React.js.
    b. Strong integration between Python Objects and JSON ensures that the backend and frontend can be tightly coupled.
    c. React.js also ensures an easy addition of features in the future. Since the components are logically connected and are structurally separated, it makes any future improvements easy to achieve.

Next, we will look at the different parts of the Application and understand the workflow and what goes on in the backend.

Application Structure and Workflow

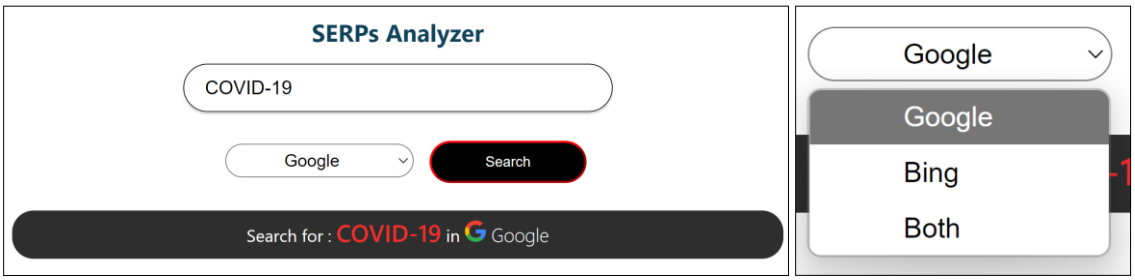


*FIGURE 2: **LANDING PAGE FOR THE APPLICATION***

On the landing page of the Application, there is a search box and a dropdown menu. The search box can be used to search for any keyword, and the dropdown menu helps choose the search engine. The 'Both' option includes a side-by-side comparison of the search results from Google and Bing.

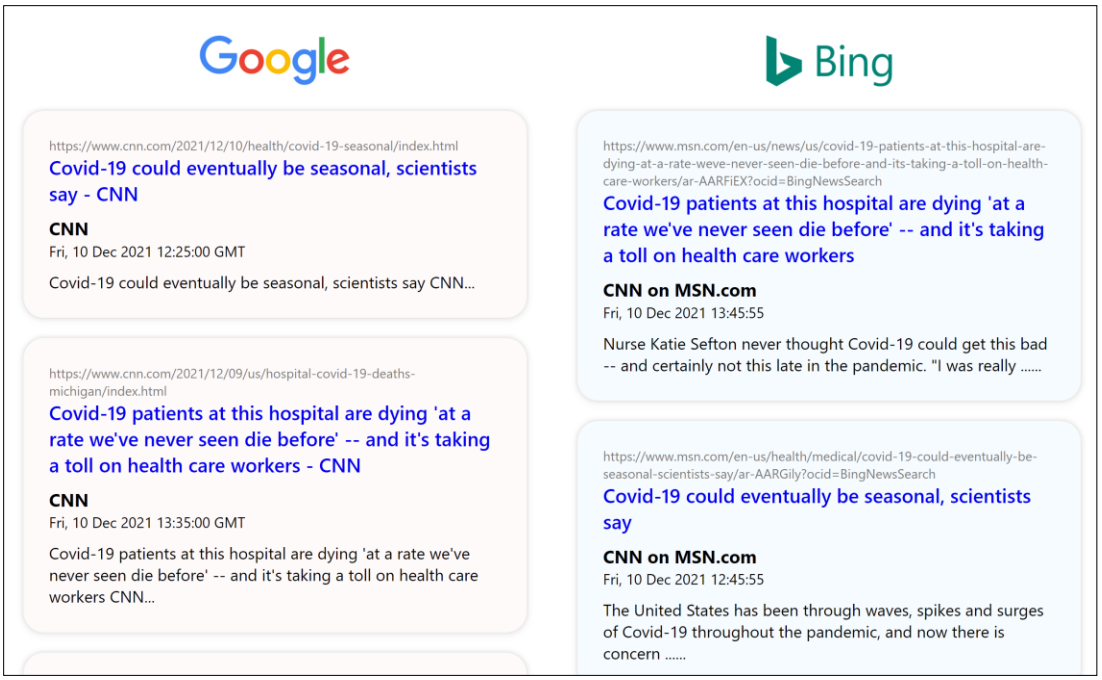


*FIGURE 3: **EXAMPLE PAGE COMPARING RESULTS FROM GOOGLE AND BING SIDE-BY-SIDE.***

When a search is made, it sends the request to the backend in the form:

```
<Top-Level Domain>/results/<keyword>/<search engine name>
```

So, in this case, it would look like this:

```
localhost:5000/results/COVID-19/both
```

This scrapes the top 10 results from Google and Bing news and formats them in a visually appealing way.

When hovering over each of the search results, it shows two buttons: one which takes to the publisher's website, and the other takes to the analysis page, which looks as follows:

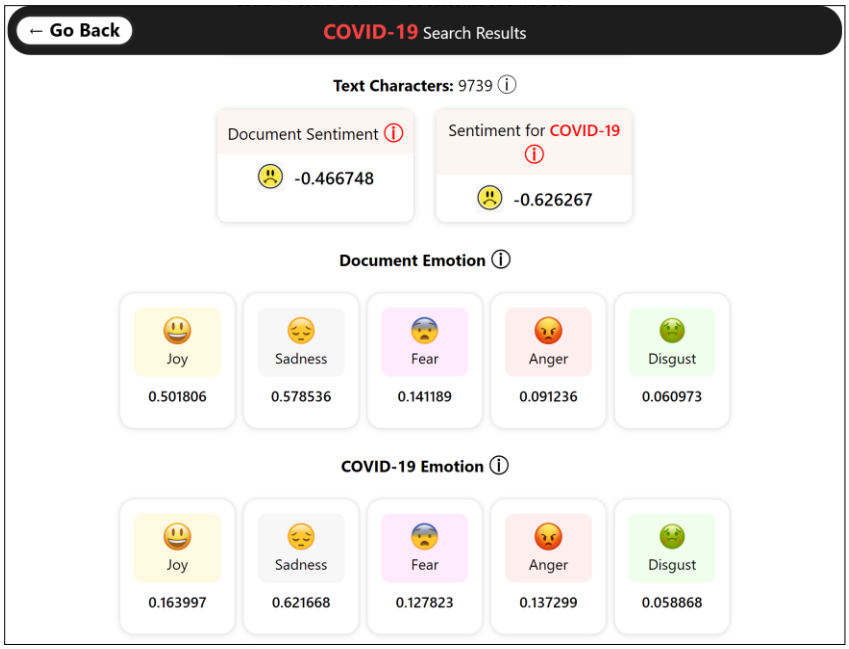


*FIGURE 4: **THE ANALYSIS DASHBOARD SHOWING THE SENTIMENT AND EMOTIONS FOR THE ARTICLE.***

The tooltip provides additional information on what each section means. As you scroll, the following sections should be present:

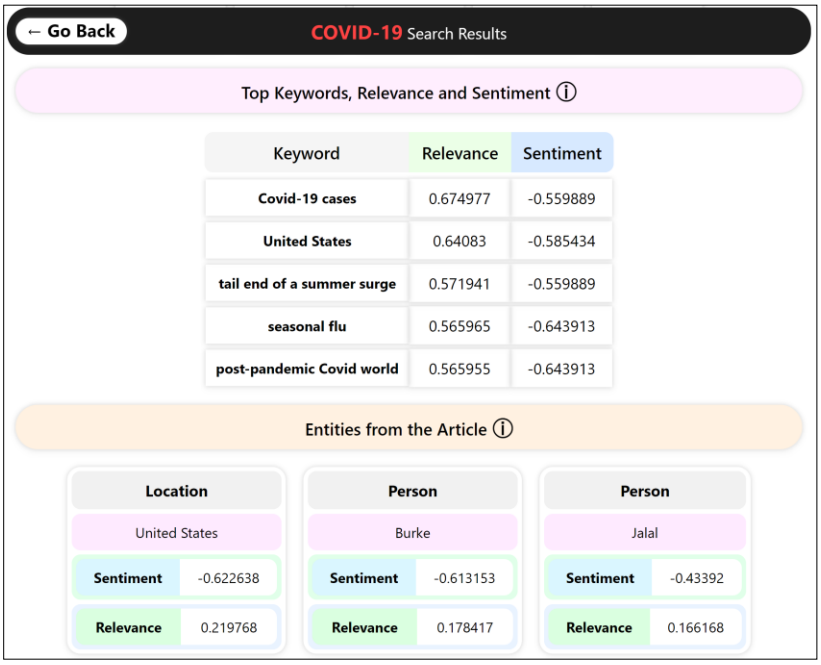


*FIGURE 5: **TOP KEYWORDS AND ENTITIES FROM THE ARTICLE***

When the 'Analysis' button is clicked, it sends a request to the endpoint in the form:

```
<Top-level Domain>/sentiment/<url>
```

This retrieves the response from the IBM servers and presents the information in an easily navigable dashboard.

Feedback and Iteration

The ultimate goal of the project is to foster an ongoing debate and create a tool to aid academic and hobbyist researchers trying to navigate the complex environment of digital media bias. So, after the first complete iteration of the data and analytics platform, I contacted 13 post-Graduate researchers across the University of Wisconsin (Madison), University of Michigan (Ann Harbor), and University of Utah (Salt Lake City) to test out the dataset, but more importantly to give feedback on the operation of the analytical pipeline.

The feedback on the intention behind the project and the parameters collected on the existing dataset was overwhelmingly positive. According to Neerajan Paudel, a Post-Grad Researcher in Psychology from UMich, the data collected alone can help test out multiple hypotheses in fields ranging from psychology to political science. Beyond that, the data analysis platform is also very robust and covers the most desirable attributes that are sought out by a researcher and provides easy access to create new graphs and other visualization aids as an exploratory tool. For example: during a demo, within a matter of a few minutes, we could obtain a table that presented the sentiment of the keyword "Kashmir" across the span of 5 months. At a high-level view, we could observe a severe increase in the sentiment value from an average of -0.37 to -0.58 before March to nearly -0.12 during March. With some context, it is clear that the release of a controversial movie demonstrating the exodus of Kashmiri Pandits in Kashmir was generating widespread recognition and appreciation from the international community. In many ways, this data is capable of drawing parallels and creating models on national and international topics and events within a matter of minutes.

Apart from the positive response, there were two main areas of improvement:

1. Beyond the applicability of the existing dataset in tackling research questions, there is a need for extensive customization of the pipeline, from the models that are used to the frequency of the data-collection process.
2. For a researcher wanting to start the data-collection process, it is critical to have a platform to monitor the operations and customize it on an ongoing basis.

Upon receiving this feedback, the platform underwent a huge overhaul in its operations and the data collection pipeline. And based on this, a backend system and a frontend component were built not just for customization but for complete automation of the process and multithreading functionality, producing a 10x reduction in the overall runtime.

Change of Database

The first major overhaul was in the infrastructure used in data storage. The move away from the NoSQL database (Google Firestore) to MySQL during the first iteration allowed for faster queries and complex joins across the tables. However, the ability to host data on cloud services would allow for a much more scalable architecture and integration of Identity Access Management (IAM) tools. So, the platform transited from MySQL to Azure SQL, an architecture based on SQL Server by Microsoft. Hosted on Azure, this system would not just allow for easy access management but would integrate well with triggers and functions native to Azure and other major Cloud platforms. On top of that, the data could be simultaneously accessed from anywhere in the world, a functionality that was missing in the local hosting of the data.

```
    server = azure_svname
    database = 'search_analysis'
    username = azure_id
    password = azure
    driver= '{SQL Server}'

    withpyodbc.connect('DRIVER='+driver+';SERVER=tcp:'+server+';PORT=1433;DATABAS
E='+database+';UID='+username+';PWD='+ password) as conn:
        with conn.cursor() as cursor:
```

```
            create_table_keyword(cursor)
```

With this transition, the user can integrate with any database through a 'pyodbc' connector just by providing the name of the server, the username, and the password to the database. This allows for any third-party integration with Cloud Providers and local hosting for a secure and personalized table schema and data patterns.

Runtime Improvements[5]

One major drawback of the system was that data could not be accessed in real-time and instead relied on batch processing. As an example, the fetches to the API for sentiment analysis and search engines across 30 keywords with ten results would take approximately 15 minutes (900 data points). This number would increase considerably if more results or keywords were added.

So, with the realization that a large portion of these operations were I/O bound, the platform embraced a multithreading architecture that would be scalable as a function of the number of keywords and number of results. Depending on the number of active threads allowed, all the operations can be completed in close to linear time since a separate background thread is opened for each set of results. The same task as before was achieved in less than a minute after this integration, which is a significant performance gain. This will now allow for an exponentially greater number of keywords and results to be collected simultaneously and with much faster runtimes.

Scheduling Tasks

With the understanding that data might be collected at different intervals, a new system based on Python's APScheduler was developed. This would allow for multiple jobs to be run simultaneously with different parameters, and the data collection could also be customized for each job. Initially, the data trigger function was running on Azure Functions (**implementation**), however, this was not scalable since most of the Cloud Platforms had a 5 to 7 minute time limit on the continuous script running, which would not be suitable for larger data collection operations.

[5] #optimization Pg. 59

And based on the interactions with academic professionals, I discovered that researchers oftentimes had access to their own servers for running such tasks. So, on top of azure functions, a custom solution was developed targeted toward such audiences. For larger data collections, a simple script could be run that would allow for highly customizable triggers and CRON jobs to be operated.

Extension of the Platform[6]

One major change during this iteration was the development of a backend and frontend that would allow for multiple jobs to be run on customized keywords and data parameters.

The platform is an extension of the existing web application and continues its development on Flask and React.js. The functional requirements of the platforms are:

1. Allow the user to choose the keywords they want to collect data on.
2. Allow the user to choose the intervals and number of results to be collected.
3. Allow the user to run queries on multiple independent lists of keyword sets with different parameters.
4. Allow the user to add, monitor, and remove jobs manually when needed.

[6] #designthinking Page. 57

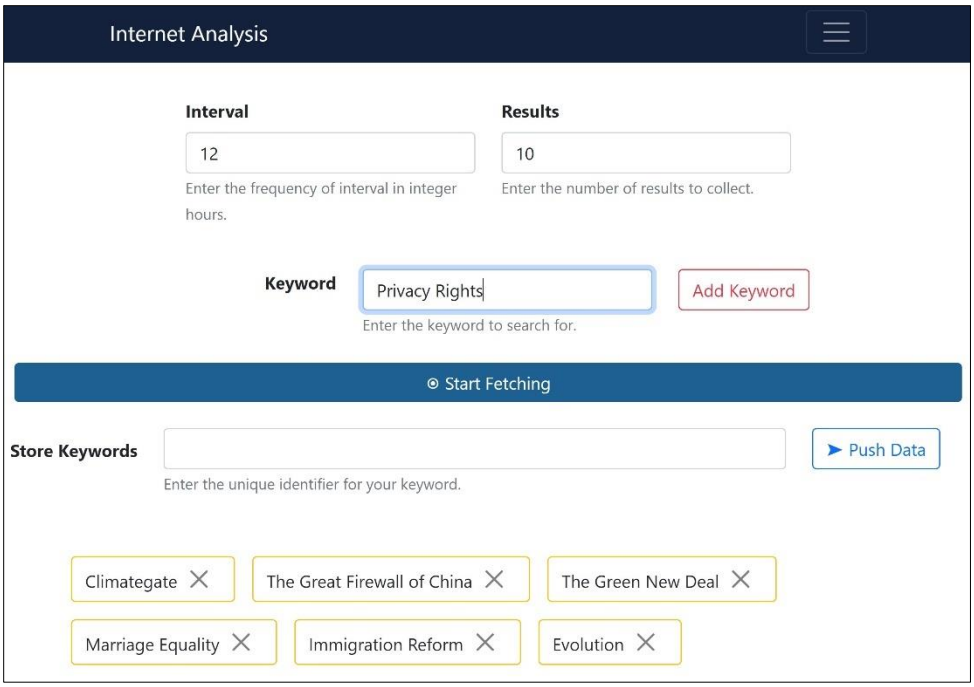


FIGURE 6: INSERTION DASHBOARD

In the Application, the user can create a list of keywords to collect the data for. Once the list of keywords is finalized, it can be stored and persisted on the SQL Server with a unique keyword. This keyword can then be used to access the information about the collected data, which provides a level of separation.

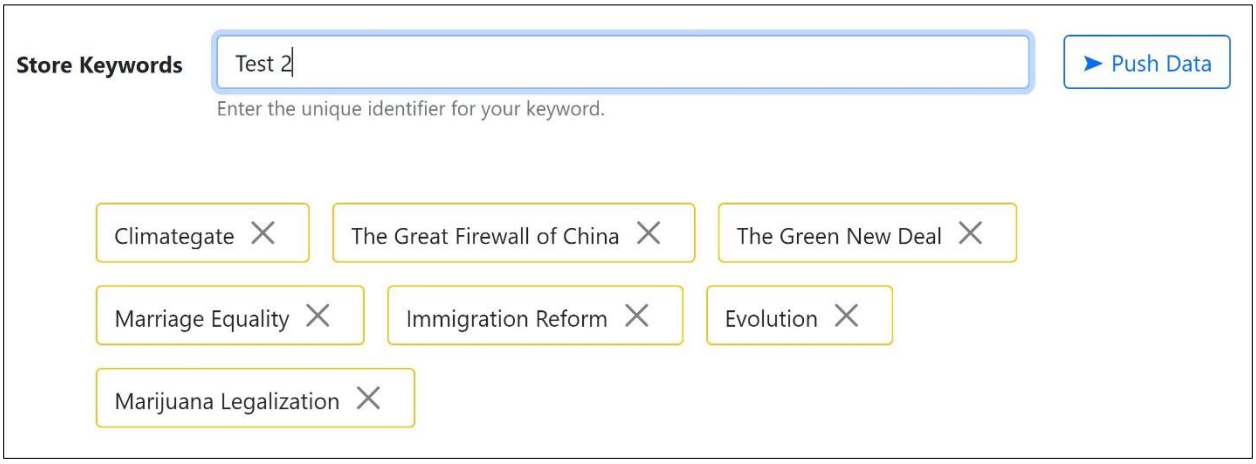


FIGURE 7: PERSISTING KEYWORDS

TABLE 2: KEYWORD IDENTIFIER TABLE

Results Messages

kw_id	kw_iden	kw_name
14	Test 1	Online Censorship
15	Test 1	Gun Control
16	Test 2	Climategate

The table stores the unique identifiers and the keyword it relates to, allowing for a one-to-many relationship.

The user also needs to specify the interval in which the data needs to be collected, along with the number of results to be collected for each platform. This allows flexibility for any customization levels and parameters.

Once this data is filled up, the user can hit the "Start Fetching" button. After this, the APScheduler schedules a job with the given parameters, and it keeps on running and collecting data over the provided time intervals. Multiple jobs can be run simultaneously through this method, as the scheduler keeps running in the background and does not block the main thread.

Another critical aspect of this is to access and delete the running jobs, as these jobs might run indefinitely at the given intervals depending on the use case. To view this, another dashboard is built.

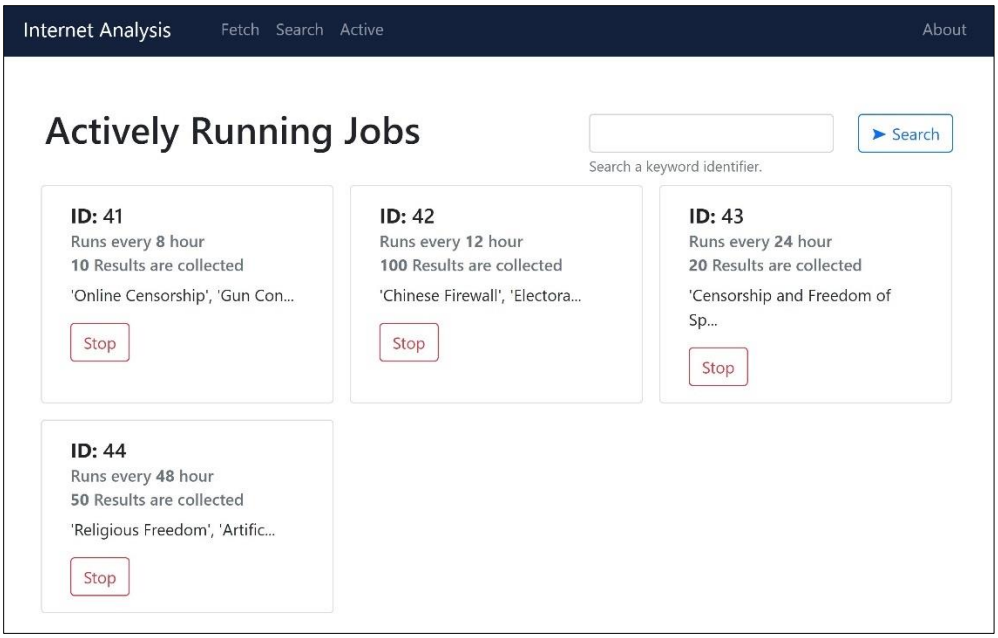


FIGURE 8: RUNNING JOBS DASHBOARD

Through the "Active" tab in the dashboard, the user can monitor all the actively running jobs and search for particular jobs by using the unique keyword identifier. From this, it is easy to monitor the details of each of the jobs and stop them if needed. This greatly facilitates the monitoring system as schedulers are oftentimes overtaxed due to resource utilization by inactive processes.

The way this works is by storing the job information in the database and tracking its current value.

Results | Messages

job_id	job_status	job_list	job_interval	job_results	kw_iden
41	True	['Online Censorship', 'G...	8	10	Test 1
42	True	['Chinese Firewall', 'Elec...	12	100	Test 3
43	True	['Censorship and Freed...	24	20	Test 4

By tracking the 'job_status,' we can check if a particular job is running and fetch all of its metadata. Each time a new job is created and run, a new row is appended to this database, which also allows for versioning. When the destructor is called by stopping the job, the 'job_status' turns into a False, and the APScheduler starts a shutdown procedure. The search function is achieved by performing a query on the 'kw_iden' column.

These additional functionalities also introduce some extra methods and endpoint APIs:

API Endpoint	GET	POST/UPDATE
`/fetchdata/`	*NONE*	Inserts the list of keywords, intervals, number of results, and the keyword identifier to the database and starts the fetch job. **Returns:** `{ 'message': 'Process has started.'}, 201`
`/activeschedules/`	Returns all the active instances of a job that are running. Fetches all the job information and metadata where "job_status" is True.	Takes in the "job_id" and shuts down the fetch process for the job. Converts the job_status from True to False and persists in the database. **Returns:** `{ 'message': 'Schedule Deleted.' }, 201`
`/searchschedules/<string:keyword>/`	Takes in the unique keyword identifier as an argument in the endpoint and searches if the keyword identifier is present and if there is a job associated with it. If there is, it returns the information about the job.	*NONE*
`/kwpersist/`	*NONE*	Takes in the list of keywords and a unique identifier and stores the information in the database.

		Returns: `{ 'message': 'Stored in Database'}, 201`

In a conversation with a political science researcher from the University of Utah, Salt Lake City, I was offered the opportunity to use the faculty servers to run larger sets of keywords and fetch a larger number of results. The runtime improvement combined with the automation allowed me to get the system up and running in a very short time. Currently, the data is being fetched for the top 30 results across Google, Bing, and Twitter across a set of 1200 keywords at intervals of 12 hours. These keywords pertain to political topics and include data to study the role of NLP techniques used by these platforms as well. A major aspect of this is the framing of the query. For example: "Does immigration harm the economy?", "Why does immigration harm the economy," "immigration and economy" all represent a query that tries to deduce the relationship between immigration and the economy, however, they are framed in completely different ways. As more data from these keywords are obtained, we can understand the patterns of sentiment and political differences across the query sets. With this, in the future, the analysis can go beyond comparative differences across platforms and move towards analyzing the nature of search algorithms and what kind of attributes NLP has in the ranking and presenting of the results.

MAKING SENSE OF DATA

DATA ANALYSIS AND FINDINGS

Analytics Pipeline

Sentimental Differences

Emotional Differences

Top Publishers

Political Leanings

Analysis of Data

With the development of the platform, it now becomes much easier to feed in new keywords and access and analyze the data. Some sample analysis is hereby presented.

Total Characters of the articles

Let's first start off by looking at a straightforward data point: the total number of characters for each of the platforms.

*TABLE 3: **AVERAGE NUMBER OF CHARACTERS AND WORDS FOR GOOGLE, BING, AND TWITTER.***

Google	Bing	Twitter
5267.9552 (740-1300 words)	4568.9649 (640-1130 words)	219.2766 (30-50 words)

From a simple metric like word count, there are many inferences we can make. On average, the articles from Google are longer than articles from Bing by around 100 – 170 words. This means with an average reading speed of 130 wpm, articles from Google take anywhere from 5.7 minutes to 10 minutes to read, whereas those from Bing take around 4.9 minutes to 8.8 minutes to read. There can be many consequences of such differences, which can be further explored using this metric.

We can also look at the correlation between character count and sentiment to see whether longer articles are more or less negatively framed. This showcases that even with a simple metric such as the total character count, there is a substantial number of inferences we can make. Breaking this further by issue types or keywords, date, etc., we can get even more nuanced conclusions.

Sentiment and Emotion across Google, Bing, and Twitter.

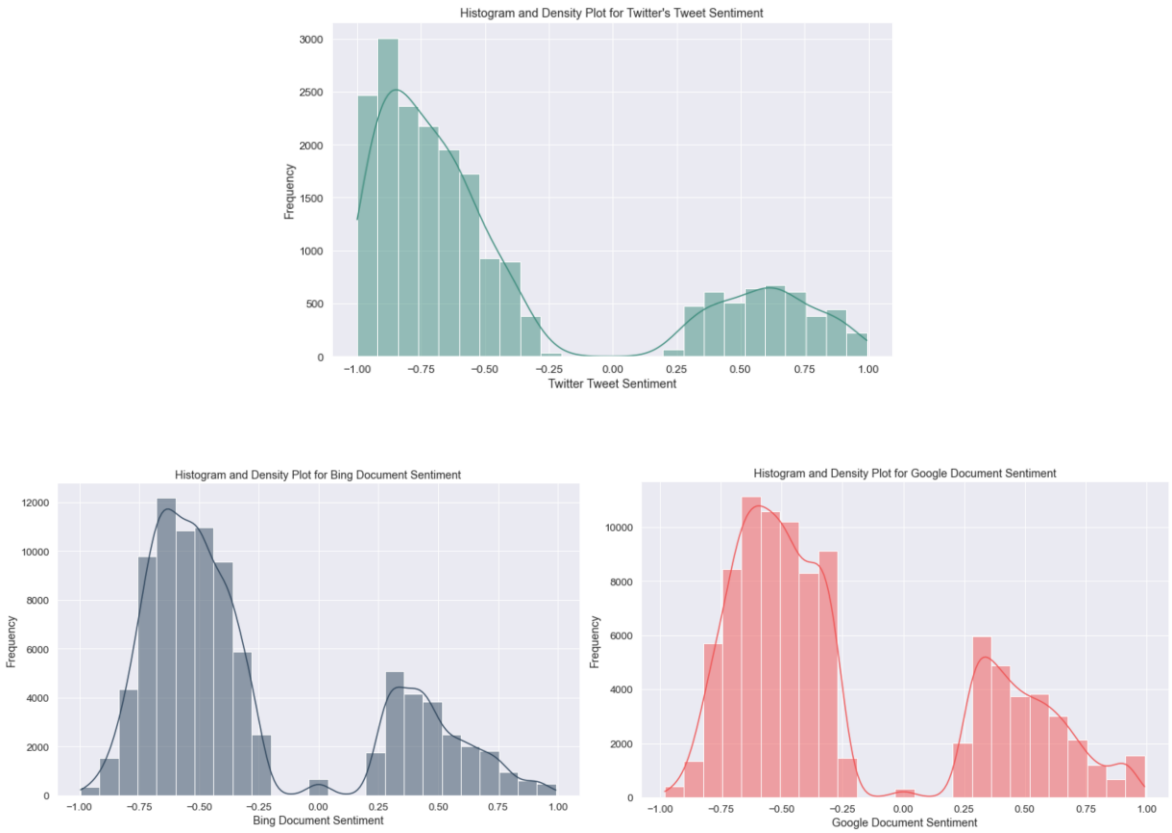


FIGURE 9: ***HISTOGRAM SHOWCASING THE SENTIMENT OF ARTICLES AND TWEETS (TOP: TWITTER, BOTTOM: BING, GOOGLE)***

Based on the histograms and density plot, we can see a clear trend in the skewness of the data. For all three platforms, there is a high volume of articles that have negative sentiment, than those with a positive sentiment, and it is distributed around the neutral sentiment, which lacks any substantial data points. The trends for Google and Bing are quite similar, but for Twitter, there is a much larger skewness towards the negative sentiments.

TABLE 4: ***AVERAGE SENTIMENT FOR GOOGLE, BING, AND TWITTER.***[10]

Google	Bing	Twitter
-0.2204	-0.2888	-0.4303

On average, we can see that the sentiment for Twitter tilts much more towards the negative side than the other two search engines. Google has the highest sentiment value, which, however, still is negative.

From the histogram, it is clear that the sentiment of the documents from all three platforms is of bimodal distribution with heavy skewness around larger values of the negative and positive sentiments, making mean an unreliable measure. So, for this reason, we also need to look at the median of the data.

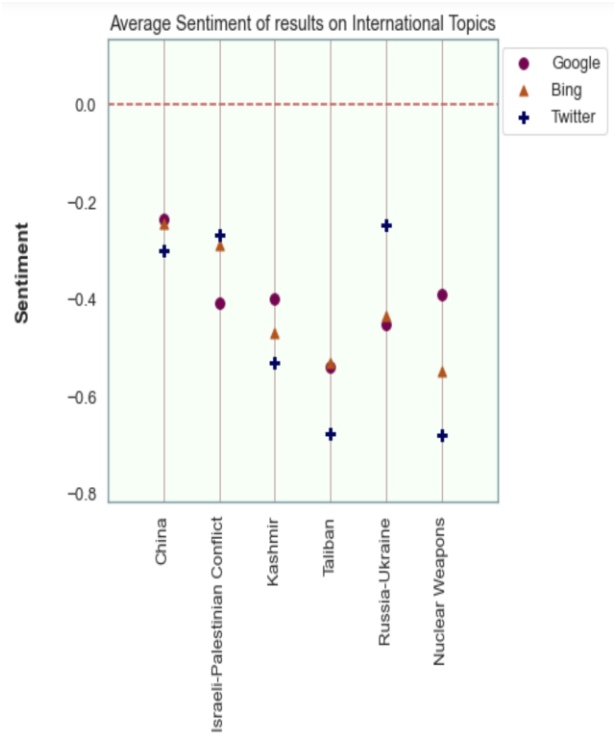
Google	Bing	Twitter
-0.4232	-0.4779	-0.669

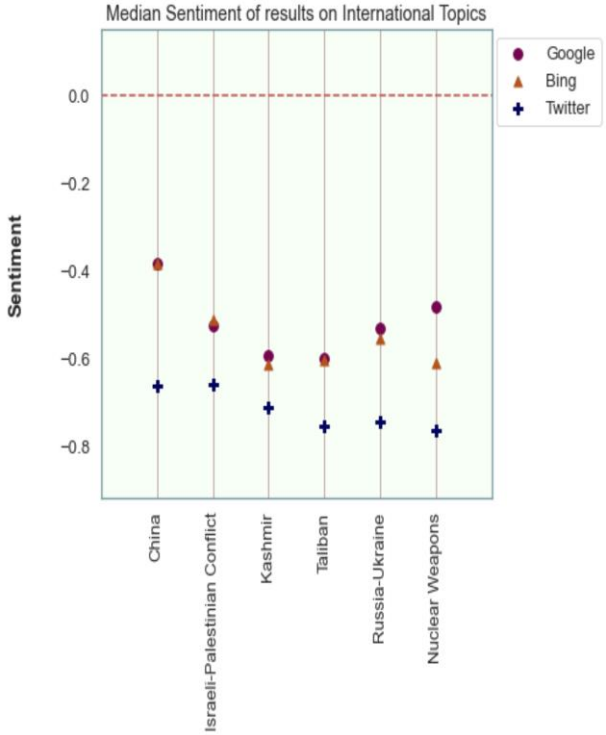
It is clear that the skewness is much more pronounced towards the negative sentiment, as there is a substantial decrease in the median across all three platforms.

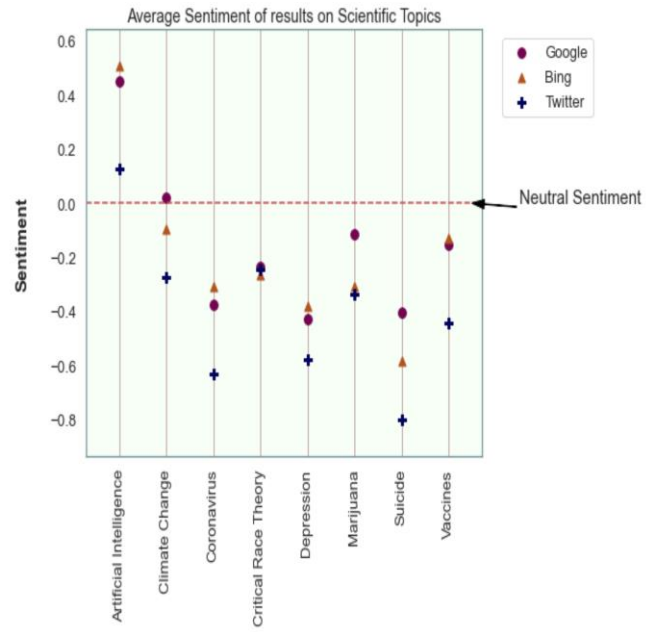
Now, let's look at trends we can see when we distribute the sentiment by issue type.

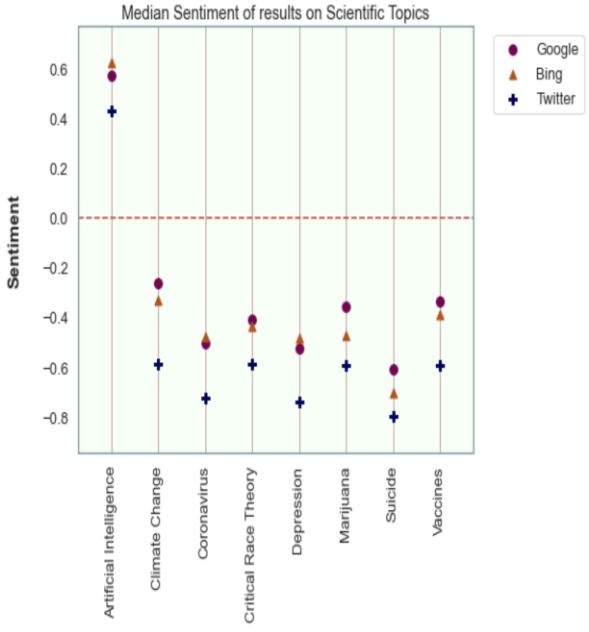
TABLE 5: ***AVERAGE SENTIMENT ACROSS PLATFORMS AND ISSUE TYPE.***

Issue/Platform	**Google**	**Bing**	**Twitter**
International	-0.4036	-0.4189	-0.4506
Social	-0.1981	-0.2740	-0.3672
Scientific	-0.1543	-0.1938	-0.3977


Average Sentiment of results on International Topics
Google
Bing
Twitter
Sentiment
0.0
−0.2
−0.4
−0.6
−0.8
China
Israeli-Palestinian Conflict
Kashmir
Taliban
Russia-Ukraine
Nuclear Weapons


Median Sentiment of results on International Topics
Google
Bing
Twitter
Sentiment
0.0
−0.2
−0.4
−0.6
−0.8
China
Israeli-Palestinian Conflict
Kashmir
Taliban
Russia-Ukraine
Nuclear Weapons


Average Sentiment of results on Scientific Topics
Google
Bing
Twitter
Neutral Sentiment
Sentiment
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
Artificial Intelligence
Climate Change
Coronavirus
Critical Race Theory
Depression
Marijuana
Suicide
Vaccines


Median Sentiment of results on Scientific Topics
Google
Bing
Twitter
Sentiment
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
Artificial Intelligence
Climate Change
Coronavirus
Critical Race Theory
Depression
Marijuana
Suicide
Vaccines

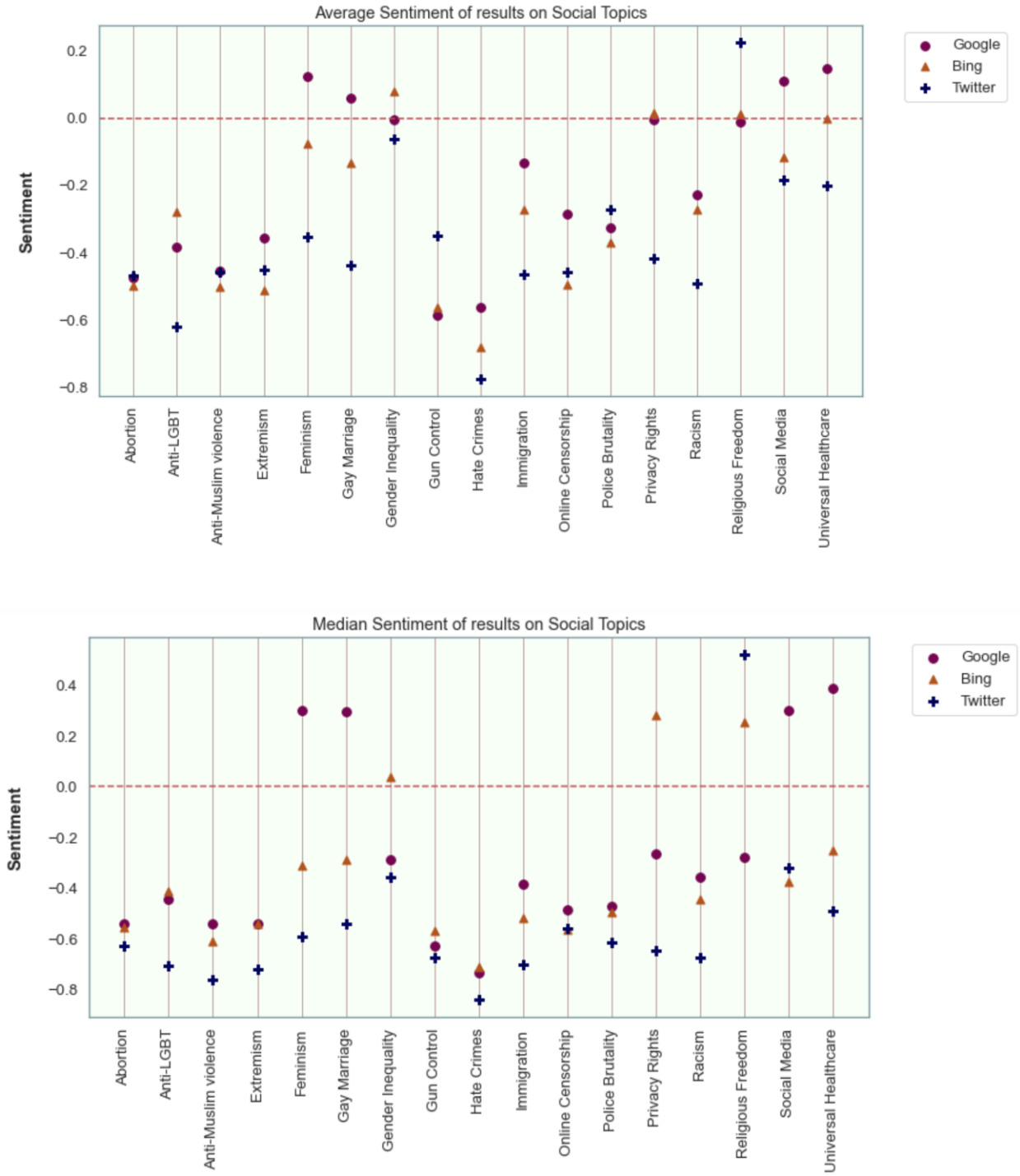


FIGURE 10: ***AVERAGE AND MEDIAN SENTIMENT OF TOPICS ACROSS INTERNATIONAL, SOCIAL, AND SCIENTIFIC ISSUES SHOWCASED IN A SCATTER PLOT.***

From the table and the plots, we can see that across all the issues, the average and median sentiment value does not change the trend seen in the sentiment for overall keywords. Google still has the highest sentiment value, and Twitter the lowest. But we can see that the sentiment toward international issues is much more negative than those toward scientific issues.

*TABLE 6: **KEYWORD SENTIMENT FOR GOOGLE, BING, AND TWITTER FOR INTERNATIONAL ISSUES.***

Sentiment on International Topics

	Google	Bing	Twitter
China	-0.235000	-0.244000	-0.301000
Israeli-Palestinian Conflict	-0.408000	-0.287000	-0.269000
Kashmir	-0.398000	-0.470000	-0.530000
Taliban	-0.539000	-0.530000	-0.677000
Russia-Ukraine	-0.451000	-0.435000	-0.247000
Nuclear Weapons	-0.391000	-0.548000	-0.680000

If we look at the keyword-wise sentiment, there is also a lot that can be noticed. For example, Google has much more negative sentiment toward inter-country conflicts than toward civil conflicts. This, singlehandedly, can be a topic for further exploration.

Beyond sentiment, we can also look at emotions and recognize the differences across those.

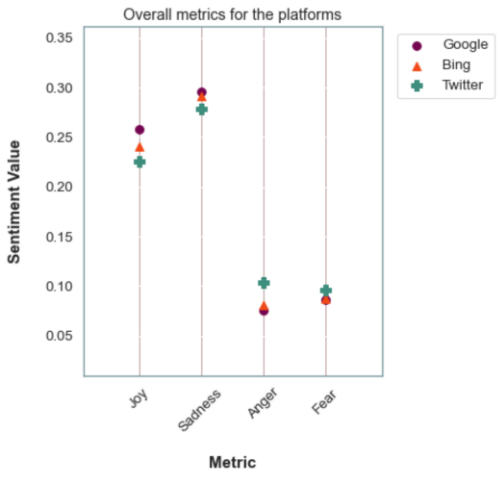


*FIGURE 11: **PLOT SHOWING EMOTIONS ACROSS THE ARTICLES***

From this, we can see that Google shows the highest magnitude of Joy and Sadness across its articles, whereas Twitter shows the highest magnitude o Anger and Fear across its tweets. Further analysis can include breaking this down by keywords or by issue types.

Rather than looking at sentiment for the overall document, we can look at the sentiment towards the entities recognized from the articles. There are multiple entity types, but the three major ones are:

1. **Person**

*TABLE 7: **SAMPLE DATA OF IDENTIFIED ENTITIES OF TYPE PERSON.***

	entity_1_type	entity_1_name	entity_1_sentiment
▶	Person	Oliver Dowden	0.43056
	Person	Quintez Brown	-0.68079
	Person	DeSantis	-0.95858
	Person	jimmy	-0.91243
	Person	Gov. Gavin Newsom	0.57017
	Person	Justin Trudeau	-0.70014
	Person	Trudeau	-0.78819

2. **Organization**

*TABLE 8: **SAMPLE DATA OF IDENTIFIED ENTITIES OF TYPE ORGANIZATION***

	entity_1_type	entity_1_name	entity_1_sentiment
	Organization	Democratic Party	-0.94273
	Organization	K-Tel	0.71369
	Organization	☑ Brings	0.98581
	Organization	Biden Administration	-0.65867
	Organization	Biden Administration	-0.77934
	Organization	NIH	-0.87584
	Organization	UNSC	0.73802

3. **Location**

*TABLE 9: **SAMPLE DATA OF IDENTIFIED ENTITIES OF TYPE LOCATION.***

	entity_1_type	entity_1_name	entity_1_sentiment
	Location	Kashmir	-0.97305
	Location	Ukraine	-0.48702
	Location	Ukraine	-0.51819
	Location	Mississippi	-0.82983
	Location	Afghan	-0.79112
	Location	Ukraine	-0.89378
	Location	Ukraine	0.55484
	Location	Canada	-0.97845

TABLE 10: ***SENTIMENT FOR THE VARIOUS GROUP OF ENTITIES***

Entity and Sentiment

	Google	Bing	Twitter
Person	-0.262240	-0.260235	-0.264330
Organization	-0.029421	-0.080661	-0.229408
Location	-0.316857	-0.298161	-0.224187

From this table, we can see that Twitter has a more negative sentiment toward people and organizations, and Google has a higher negative sentiment toward locations.

Beyond this analysis of sentiment and emotions, we can look at the nature of publishers that the search engines represent.

Publisher frequency and sentiment

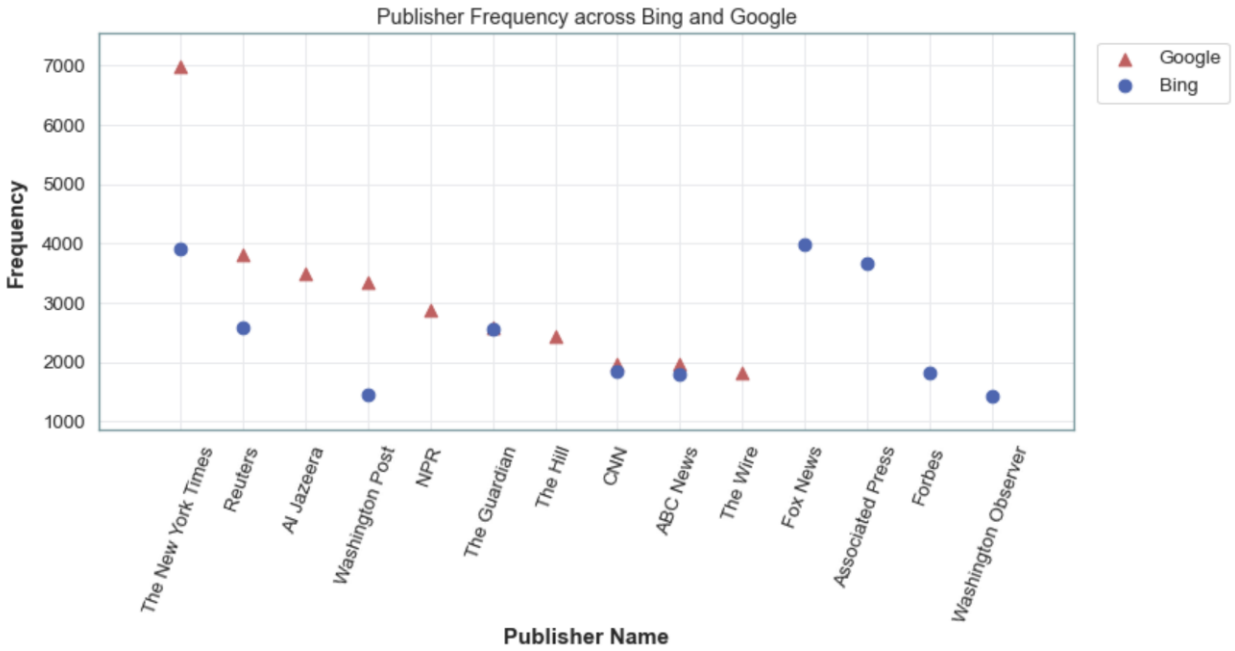


FIGURE 12: ***SCATTER PLOT TO SHOW THE TOP 10 AND THEIR FREQUENCIES.***

We can see six common publishers in the top 10 frequencies list between Google and Bing, whereas the rest differ.

*TABLE 11: **PARTISANSHIP OF THE PUBLISHES AS DETERMINED BY ALLSIDES.***

	Google	Bing
center	3.0	3.0
left	1.0	0.0
left-center	6.0	5.0
right-center	0.0	2.0

For a more nuanced look, we can even see the partisanship of the top publishers for Google and Bing. There are no right-leaning publishers that are in the top 10 list for Google, whereas Bing has a more even mix of publishers.

We can further look into a more visual representation of these data.

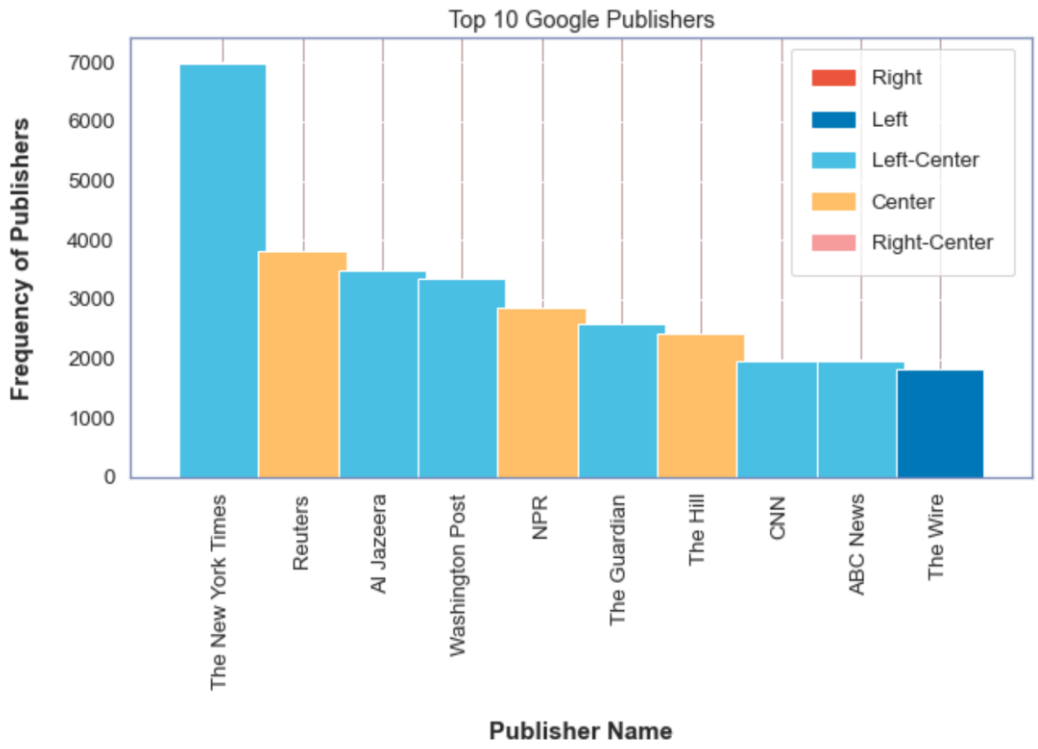


FIGURE 13: ***PUBLISHER FREQUENCY AND POLITICAL LEANING ACROSS GOOGLE AND BING.***

There are many conclusions that can be made from this data regarding the inclusion of differing viewpoints on these platforms.

Not only can we derive insights into the search engines, but we can also look closely at the publishers.

From the diagram below, we can see what the average sentiment for each of these publishers is. However, this can be by virtue of the topics they cover, so we can even divide them into different issues and even keywords.

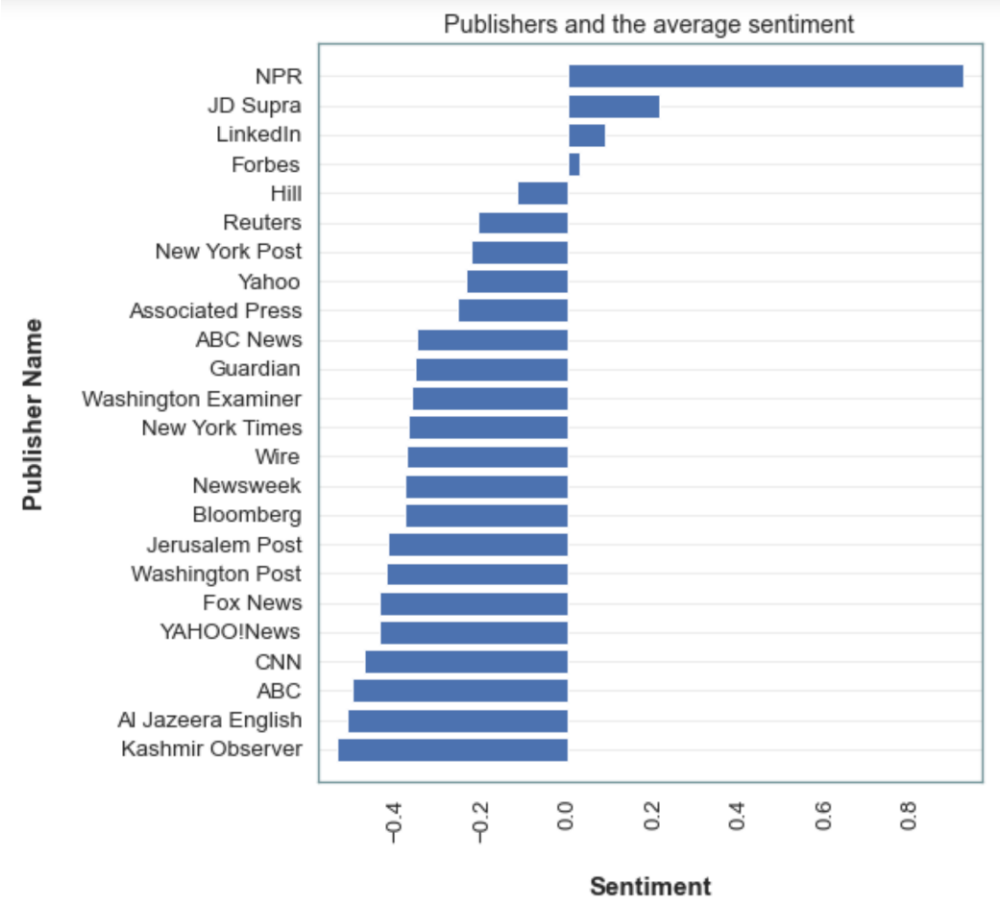


FIGURE 14: ***THE TOP PUBLISHERS AND THEIR AVERAGE SENTIMENT.***

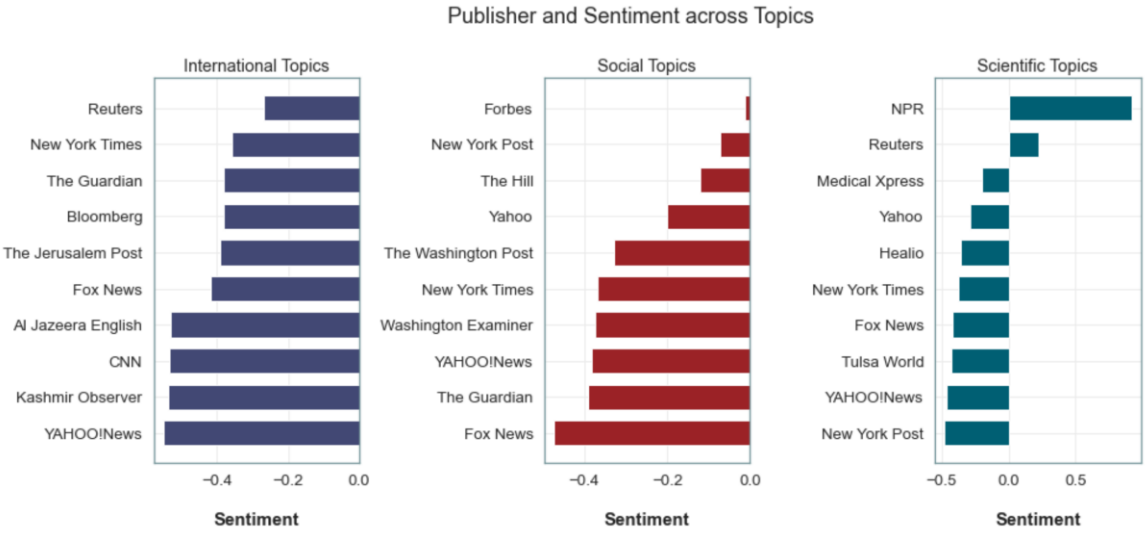


FIGURE 15: ***SENTIMENT OF TOP PUBLISHERS BY ISSUE TYPE***

From the diagram, we can see the sentiment of the top publishers across each of the three domains. We can even compare the sentiment of the same publisher across these domains. For example, Yahoo! News has a much more negative sentiment in the articles covering international topics than in those covering social topics. This trend is exactly the opposite for Fox News.

Through the same set of data, not only can we analyze the various platforms but also the various publishers. And the rate at which people view these articles is very well represented by top search engine results since a majority of the people use search engines as a gateway to their news and also click on the top few results of the results page.

This is only a glimpse of the kind of insights we can derive by analyzing the unobvious aspects of news articles and these platforms. There are many directions that the analysis can head towards. From analyzing the correlation between partisanship and sentiment to that between retweets and favorites count of twitter on the emotions. A single aspect of these metrics can guide many research projects, and much more nuanced insights can be gathered as the database grows in data points. In the future, building even larger datasets covering thousands of keywords, and more targeted selection of topics will pave the path towards a better understanding of the impact of internet platforms on our society. This project marks a significant step towards that end.

BUILDING UPON THE SHOULDER OF GIANTS

REFERENCES

References

Adams, S. (2016, February 19). *The Founder of DuckDuckGo Explains Why Challenging Google Isn't Insane*. Retrieved from Forbes: https://www.forbes.com/sites/forbestreptalks/2016/02/19/the-founder-of-duckduckgo-explains-how-to-get-customers-before-you-have-a-product-and-why-challenging-google-isnt-insane/?sh=63e0db734e89

Allsides. (n.d.). *Allsides*. Retrieved from Allsides: https://www.allsides.com/unbiased-balanced-news

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. B., . . . Volfovsky, A. (2018, August 09). *Exposure to opposing views on social media can*. Retrieved from PNAS: https://www.pnas.org/content/pnas/115/37/9216.full.pdf

Barberá, P. (2014). How Social Media Reduces Mass Political Polarization. Evidence from Germany, Spain, and the US.

Boutin, P. (2011, May 20). *Your Results May Vary*. Retrieved from The Wall Street Journal: https://www.wsj.com/articles/SB10001424052748703421204576327414266287254

Boyd, D. (2015, May 11). *Social Media: A Phenomenon to be Analyzed*. Retrieved from SAGE Journals: https://journals.sagepub.com/doi/10.1177/2056305115580148

Brin, S., & Page, L. (2003). The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Science Department, Stanford University, Stanford, CA 94305*.

Caren, J. (2021). *The Modern, Inside Scoop on Google PageRank In 2021*. Retrieved from Hubspot: https://blog.hubspot.com/blog/tabid/6307/bid/45/the-importance-of-google-pagerank-a-guide-for-small-business-executives.aspx

Carl Hendy. (n.d.). *The History of Search Engines*. Retrieved from Carl Hendy: https://carlhendy.com/history-of-search-engines/#archie

Chen, W., Pacheco, D., Yang, K.-C., & Menczer, F. (2021, September 22). *Neutral bots probe political bias on social media*. Retrieved from Nature Communications: https://www.nature.com/articles/s41467-021-25738-6

Cherry, K. (2020, April 29). *What Is the Negativity Bias?* Retrieved from VeryWellMind: https://www.verywellmind.com/negative-bias-4589618

Ciampaglia, G. L., Nematzadeh, A., Menczer, F., & Flammini, A. (2018, October 29). *How algorithmic popularity bias hinders or promotes quality*. Retrieved from Scientific Reports: https://www.nature.com/articles/s41598-018-34203-2

Cornell University. (2018, October 18). *Bing and Google: How are the Search Algorithms Different?* Retrieved from https://blogs.cornell.edu/info2040/2018/10/18/bing-and-google-how-are-the-search-algorithms-different/

Couvering, E. V. (2008, January). *The History of the Internet Search Engine: Navigational Media and the Traffic Commodity*. Retrieved from Springer Link: https://www.researchgate.net/publication/227282790_The_History_of_the_Internet_Search_Engine_Navigational_Media_and_the_Traffic_Commodity

Dean, B. (2020, August 20). *How People Use Google Search*. Retrieved from Backlink: https://backlinko.com/google-user-behavior

DiFranzo, D., & Gloria-Garcia, K. (2017, April 5). *Filter bubbles and fake news*. Retrieved from ACM: https://dl.acm.org/doi/10.1145/3055153

Digital Synopsis. (n.d.). *Why Page 2 Of Google Search Results Is The Best Place To Hide A Dead Body*. Retrieved from Digital Synopsis: https://digitalsynopsis.com/tools/google-serp-design/

Featherly, K. (2016, May 11). *ARPANET United States defense program*. Retrieved from Britannica: https://www.britannica.com/topic/ARPANET

Garrett, R. K. (2009). Echo chambers online?: Politically motivated selective exposure among Internet news users. *Journal of Computer-Mediated Communication, Volume 14, Issue 2, 1*, 265–285.

Gleich, D. F. (2015). PageRank Beyond the Web. *SIAM Review, Vol. 57, No. 3*, 321-363. Retrieved from https://www.jstor.org/stable/pdf/24778735.pdf

Goldman, E. (2006, January 1). *Search Engine Bias and the Demise of Search*. Retrieved from Santa Clara Law: https://digitalcommons.law.scu.edu/cgi/viewcontent.cgi?article=1112&context=facpubs

Gregersen, E. (2020, February 14). *Bing search engine*. Retrieved from Britannica: https://www.britannica.com/topic/Bing-search-engine

IBM Cloud. (2021, September 09). *Natural Language Understanding*. Retrieved from IBM: https://cloud.ibm.com/apidocs/natural-language-understanding?code=python

Islam, A. N., Laato, S., Talukder, S., & Sutinenb, E. (2020, October). *Misinformation sharing and social media fatigue during COVID-19: An affordance and cognitive load perspective*. Retrieved from PMC: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7354273/

K, S., & Patil, D. R. (2013). Social Media – History and Components. *IOSR Journal of Business and Management (IOSR-JBM)*, 69-74.

Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014, June 17). *Experimental evidence of massive-scale emotional contagion through social networks*. Retrieved from PNAS: https://www.pnas.org/content/111/24/8788

MacArthur, A. (2020, November 25). *The Real History of Twitter, in Brief*. Retrieved from Lifewire: https://www.lifewire.com/history-of-twitter-3288854

Meyer, J. (2020, January 2). *History of Twitter: Jack Dorsey and The Social Media Giant*. Retrieved from The Street: https://www.thestreet.com/technology/history-of-twitter-facts-what-s-happening-in-2019-14995056

Nexor. (2014, August). *ALIWEB: THE WORLD'S FIRST INTERNET SEARCH ENGINE*. Retrieved from Nexor: https://www.nexor.com/aliweb/

Pramatarov, M. (2018, December 27). *DNS history. When and why was DNS created?* Retrieved from Cloudns: https://www.cloudns.net/blog/dns-history-creation-first/

Radd Interactive. (n.d.). *Bing's Ranking Factors & Algorithm – How to Rank on Bing*. Retrieved from Radd Interactive: https://raddinteractive.com/bings-ranking-factors-algorithm-how-to-rank-on-bing/

Ribeiro, M. H. (n.d.).

Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A., & Meira, W. (2020). Auditing radicalization pathways on YouTube. *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, (pp. 131–141).

Rosenberg, E. (2020, June 23). *How Google Makes Money (GOOG)*. Retrieved from Investopedia: https://www.investopedia.com/articles/investing/020515/business-google.asp

Science Museum. (2018, November 2). *From ARPANET to the Internet*. Retrieved from Science Museum: https://www.sciencemuseum.org.uk/objects-and-stories/arpanet-internet

Shearer, E. (2021, January 12). *More than eight-in-ten Americans get news from digital devices*. Retrieved from Pew Research Center: https://www.pewresearch.org/fact-tank/2021/01/12/more-than-eight-in-ten-americans-get-news-from-digital-devices/

Shewan, D. (2014, June 23). *How Google Hummingbird Changed the Future of Search*. Retrieved from WordStream: https://www.wordstream.com/blog/ws/2014/06/23/google-hummingbird

Smallwood, L. (2012, April 13). *Optimizing Website Content for Higher Rankings on Bing*. Retrieved from Microsoft Bing Blogs: https://blogs.bing.com/uk/2012/04/13/optimising-website-content-for-higher-rankings-on-bing

StatCounter. (2021). *Search Engine Market Share Worldwide*. Retrieved from StatCounter: https://gs.statcounter.com/search-engine-market-share

Statista. (2022). *Number of social network users worldwide from 2017 to 2025*. Retrieved from Statista Research Department: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

Web Foundation. (n.d.). *History of the Web*. Retrieved from World Wide Web Foundation: https://webfoundation.org/about/vision/history-of-the-web/

Wikipedia. (n.d.). *Random surfing model*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Random_surfing_model

Wikipedia. (n.d.). *WebCrawler*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/WebCrawler

Wikipedia. (n.d.). *Wikipedia: List of controversial issues*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues

THE IMPORTANT AND THE MISFITS

APPENDIX

Capstone Learning Outcomes
Course Learning Outcomes (LOs)
Collection of the HCs
Extra Graphs and Visual Aids

Appendix

Capstone LOs

1. **#qualitydeliverables:** I have changed the priorities of the project to ensure that it better tackles the main problem at hand. By improving the rigor of techniques and specifications of platforms used for prior analysis, I have created a much higher quality scalable system that ensures data collection beyond the scope of this Capstone. After realizing that even the smallest aspect of the data can provide insights into much deeper parts of the ecosystem, I spent time figuring out the best way to communicate this, which I do through examples across my analysis. I have also followed all the guidelines for the project and met all the requirements for it while incorporating the feedback received throughout this process.

2. **#navigation:** Starting from the humble beginnings of my Capstone to its current state, I feel much stronger about the direction it has taken. The decision to focus on creating a tool and categorizing a list of findings from it to guide future research, rather than focusing on taking a stand on a very complicated topic, has provided a novel contribution to this field. Based on my research, I haven't seen any comparable tools and data that are available on the Internet, but there are plenty of interpretations of existing data and studies with below-par metrics to support the qualitative analysis. This brings about a much-needed quantitative tool, and for this, I have carefully crafted and re-edited the purpose of this project.

3. **#metrics:** The metrics I have chosen highlight some important aspects of the project, from the process of managing and writing scalable code to target the right audience via abundant abstractions. These metrics have remained more or less the same from prior rounds since I think that strengthening these metrics is very important, and it also helps me track my progress and the proximity to the end goal. I also actively sought out feedback from the target audience as a measure of relevance and practicality of the developed solution.

4. **#curation:** I have used the appropriate medium to present the code (GitHub) and included only the necessary elements of my analysis in the main paper. The rest of the diagrams are included in the Appendix, in need of further reference. This prioritization offers a clearer and

more organized analysis, one that flows smoothly from one topic to another. The organization also moves from the context to framing the question, proposing a solution, and justifying its viability. It leads to a complete and strong narrative for my Capstone. Also, based on the feedback and guidance received during the Final Draft round, I contacted researchers across three universities and demonstrated the project to them. Based on the feedback received from these interactions, I was able to prioritize well and extend the application functionality as well.

HC Appendix

(HCs pointing to certain elements of the paper has a two-way link at the end of the paragraph, and the footnote on the paper brings back to the Appendix, smoothening the navigation process.)

1. **#carrotandstick:** PageRank is a great example of a carrot and stick mechanism that prevents negative behaviors that were prevalent in search engines of the past. Mathematically, it shows that methods to game the system, like including excessive outbound links to a website, would lead to a decline in its PageRank. And, with the introduction of "Hummingbird," only relevant articles with appropriate keywords were rewarded. There are two main points in this:

    a. Including unnecessary and irrelevant keywords were no longer rewarded by PageRank since it focused more on anchor links and text.

    b. Including unnecessary outbound links to promote the page's PageRank would also be punished since it decreased the PageRank and hence affected the PageRank of all the connected links.

    c. Incentivized high-quality content since it would be picked up by the web crawlers from Google, and natural language understanding made including keywords not crucial to get a higher PageRank, encouraging small creators to spend less time in Search Engine Optimization (SEO) and more time in creating good content.

2. **#context:** Starting from the beginning of the Internet, all the way to modern search engines, the reader can understand the context behind why search engines came into existence, what kind of problems they faced, and how that influenced the ranking system of modern search engines. There is also ample background into the origins of three prominent search engines:

Google, Bing, and DuckDuckGo, where Google and Bing are further discussed in the context of their internal mechanisms and ranking algorithms. Furthermore, I discuss why some of the search engines survived and others didn't, which might have some interesting consequences. Beyond this, I also present the context behind blogging as an information dissemination tool and the origin of Twitter.

3. **#sampling:** Based on the search engine traffic, which highlights the fact that most of the users only interact with the first few results that are presented, we deduce the sampling mechanism. By collecting the data only from the first ten results, it ensures that these are the most widely representative articles that are prioritized by the two search engines: Google and Bing. Further, the sampling is conducted in the pre-defined list of keywords, which are used to exaggerate the sentiment and emotional expressions to make the differences between the search engines more visible. Another decision was to determine which sample of platforms would be ideal for analyzing throughout the project. Based on the usability and influence of the platforms in information dissemination, the final sample chosen were Google, Bing, and Twitter.

4. **#medium:** The results are represented in the form of a Jupyter Notebook, which ensures that the user can obtain their own data and gather results on a real-time basis. This is also complemented with a web application, which provides an instant time lookup upon the articles from the two search engines, and also gathers results about the sentiment of the document content. This medium ensures not just interactivity but the collection of the most up-to-date and recent information analytics. This was further strengthened by a new dashboard to create fetch jobs and monitor them within the Application, providing a visual aid to the end-user. Plus, the project analyzes two of the most important mediums for information gathering on the Internet, search engines and social media. It provides tools to critically examine the nature and traits of these mediums and determine their suitability for various aspects.

5. **#desginthinking:** The project has undergone many substantial changes throughout its lifecycle. From the questions it tackles to the way it is presented, all these changes were made to ensure a strong target for the main thesis of the paper. The analytical methods and even the datastore went through many stages of iteration. For example: in the beginning, I used Google

Firestore to store the information, but also seeing the limited reads and writes it allows, I decided to use CSVs for the same task. However, this created many problems in querying the data, so eventually, I decided to use a SQL database, which was not only much faster but also allowed compound queries.

When designing the Application, and the analysis, it went through multiple stages of iteration and was substantially improved upon. Initially, the Application was built using the jinja template in Flask, without a separate frontend. But the templating features were extremely limited, and the logic inside the HTML pages didn't scale properly. Also, the code got very congested as there was no way to logically break down the components. It was clear at that point, a new frontend was needed, which is when I moved to React.

The final iteration was after getting the feedback from other researchers, which is when I decided to change the database and develop an application that could schedule and monitor jobs with great customizability. 6

6. **#variables:** In order to make the analysis robust, there were many variables that were collected and created. Using the document content and the keyword, we could get the sentiment behind the article, and using the publisher information, we could obtain the political leaning of them as well, all of which helped to form a strong analysis. 4

7. **#heuristics:** While looking at PageRank and Bing's ranking algorithms, we look at what heuristics they use and how the emphasis they place upon each varies, leading to different results. This is complemented by the earlier discussion of outdated heuristics, like keyword search, and how they were manipulated by web admins leading to undesirable outcomes. 1

8. **#complexcausality:** In the background context, I discuss the various factors that led to the current implementation of search engine algorithms and digital platforms. Primarily, we look at how the automated nature of search engines, along with the heuristics it utilizes can cause unexpected outcomes. For example, the removal of personalization might lead to outcomes that are undesirable, although, on the surface it seems to tackle the issue of an individual-level filter bubble. Going deeper, however, we see a behavior where only the viewpoints of the

majority would be included and would be primarily dominated by big web pages with lots of resources in SEO and marketing, a finding supported when analyzing the top publishers across Google and Bing. By understanding the internal mechanisms and incentives of the publishers of these platforms, we can understand why filter bubbles are prevalent and why some reinforcing feedback loop makes it harder to break out of it. 2

9. **#critique:** Especially in discussing filter bubbles, it is clear that the evidence of personalization on the Internet and its effects on intellectual isolation is mixed. So, we go through both sides by looking at previous studies that seem to confirm that there is a great deal of nuance on what works and what doesn't. This critique of the filter bubble idea also leads to the premise of the analysis, which is that there might be something more implicit and unobvious that might be leading to certain behaviors.

10. **#descriptivestats:** This is the foundation of exploratory analysis, which primarily consists of the descriptive stats for the data that is collected, along with the necessary data visualizations to complement it. This is also complemented by the interpretation of the results. On top of that, many further discussions and analytical methods were inspired by the unique findings from such stats and metrics. For example: the bimodal distribution of the histogram inspired the collection of median data as well as the mean.

11. **#conformity:** Filter bubbles and PageRank are discussed, along with their consequences which help explain the reinforcement of certain behaviors of automated algorithms and how the problem is compounded by human conformity. Especially with personalization, people are used to seeing materials on the Internet that cater to their existing beliefs and are less open to new ideas. This is shown especially by the study discussed in the paper, where social bots introducing opposing views to Twitter users had an opposite impact where it ended up reinforcing their beliefs instead opening up the path to a new one. 3

12. **#optimization:** To ensure that the platform is functional as a research tool, major consideration has been made in terms of the fastness and smoothness of the pipeline. As such, optimizations have been a major part of the design and iteration process. From the transition across NoSQL to SQL database for faster queries to the inclusion of multithreading and

scheduling tools, the goal is to optimize both the run time and practical operational quality of the tool. Through experimentation and practical testing with the researchers, I was able to optimize the tool and prioritize features that fit the need of the targeted use case. An example can be the reduction of overall runtime for a single fetch from 15 minutes to less than 1. 5

13. **#professionalism:** All the guidelines for the report are met, and the paper is organized in a systematic and logical manner, with appropriate language. The same holds true for the code base, which is properly documented, to a stage where readers can get the servers and systems up and running by following the steps included in the GitHub repository.

14. **#constraints:** There were many constraints that had to be considered when collecting and analyzing the data, especially for the web application. For example: due to the limited number of calls that can be made to IBM servers, instead of having a high-level summary for a keyword, the app provides a dashboard for each of the results that ensures only one call is made for a result. Even when designing the Application, instead of using repetition and distinct methods for SQL, I pre-processed the data to ensure that only necessary calls were made. As an instance, if the sentiment for the same URL had already been calculated for Google, it needn't be calculated again for Bing if the same article is encountered.

15. **#algorithms:** There are numerous algorithms that are used through the Application and the analysis, from minor algorithms to sort data using Heaps to machine learning algorithms to gather sentiment data. Routing algorithms are also used to create the RESTful API for the backend, and these algorithms are well-documented across the codebase. Beyond this, I also provide a theoretical description of the PageRank algorithm and its mechanisms, along with some extensions of the algorithm as implemented by Google and Bing.

16. **#audience:** The audience for the project is clear, i.e., academic and hobbyist researchers wanting to explore the nature of digital platforms. This audience has been the foundation during the development process, and especially during the feedback gathering and final iteration. By directly communicating with the intended audience, I was able to gather a list of requirements, prioritize them and start the development, with the end product being much more robust and practical for the researchers. Example is the added functionality to customize

the keywords and other parameters to collect independent data targeting the research question at hand, along with a job monitoring system.

17. **#induction:** The question this project is trying to tackle is only possible through induction, and the design is of such a way that it would strengthen this inductive reasoning. It is not feasible to gather and analyze all the results and keywords before concluding and presenting the patterns across these platforms. Instead, by gathering a representative set of data across time and by providing qualitative justifications of the design and keywords decisions, it is possible to extrapolate the findings and make broad generalizations that are of a practical significance. Another strength of it is that as more data is collected temporally, the induction would only get stronger as there will be higher volume of data points to base our findings upon.

18. **#purpose:** The purpose of the project, from the very beginning, was to understand what kind of biases and patterns are present in the digital platforms that might cause desirable and undesirable situations in the society. The method with which I tackle this problem has gone through several rounds of iteration, but the final product developed directly tackles the core of the issue by generating engagement in the field, and providing preliminary data and tool to showcase the empirical value behind such analysis. With the guidance of my Capstone Advisor, I was able to refine the purpose of the project and delve into the problem with much critical and detached eyes, which ultimately added value to the end product.

LO Appendix

1. **#abstraction** (CS162): The main function used ensures that all the decision-making, cleaning, and insertion of data can be done using two parameters. The rest of the methods are hidden, so the client does not have to worry about the details of how the function works.

   The Application for the project uses a React.js frontend, where abstraction is achieved by composition. Dividing each component of the user interface into smaller components and then using these lower-level components to obtain higher-level components ensures that much of the logic and complexity is hidden, and only the necessary data and functions are

exposed. For example, the Result Component only needs to know what the article title, description, date, and URL are. So, these data are exposed to the higher-level component, whereas the child component is separated in a different file altogether. Similar parallels can be drawn across each of the pages.

2. **#webstandards** (CS162): The standards for web architecture, web accessibility, and web of devices are all met by the Application. The Application is built on the latest RESTful backend using Flask, which only exposes the necessary endpoints: to receive the articles from the search engine, to receive the sentiment of those articles, to schedule tasks, to connect with the database, etc. It also uses the best practices of RESTful identifiers, and the URIs are formatted based on those requirements. Also, by using React.js, Grids, and Bootstrap to style the components, it ensures a uniform experience across devices and ensures the responsiveness of the webpages, along with accessibility concerns.

3. **#separationofconcerns** (CS162): Separation of logically and computationally similar tasks has been given a very high priority, not only to keep the codebase clean but also to ensure that other functionalities in the future can be easily added. There are three main aspects of this:
   a. The API for the Flask application is separated out into multiple endpoints that are logically connected but aren't always called simultaneously. This ensures that only the minimum number of calls are being made, especially to the IBM server since it is a pay-as-you-go plan.

   b. In the React App, the logically connected components are separated out and reused whenever needed. This prevents the repetition of code. This was also the reason why React.js was chosen instead of other templating engines since it provides a high degree of flexibility in structuring the Application.

   c. The database, schedulers, front end, and backend are developed from the perspective of a microservices architecture using a RESTful API. This allows the decoupling of a system and makes debugging much simpler, as it is easier to pinpoint the cause of failure.

4. **#attentionroles** (SS110): This LO is especially relevant in discussing filter bubbles, personalization, and the role of emotions. In the section, we discuss what kind of headlines and articles gain more attention and how the position of the search results influences the click-through rate. Based on these findings, the analysis is built in a way that only the top 10 results are considered from the search engines since the traffic reduces significantly from that point onwards. The emotion and sentiment conveyed in the articles also hint at a top-down attention mechanism, where based on the previous knowledge and intention behind the search query, certain articles seem more enticing than others, which influences the ranking mechanism of the search engines.

5. **#data** (SS154): The 'Data collection and analysis section includes all the relevant details of what data is being collected, what it signifies, and how it is being collected. By including this information and justifying their choices, the reader can follow through rest of the analysis, and it also helps strengthen the findings from it.

Other diagrams from the analysis

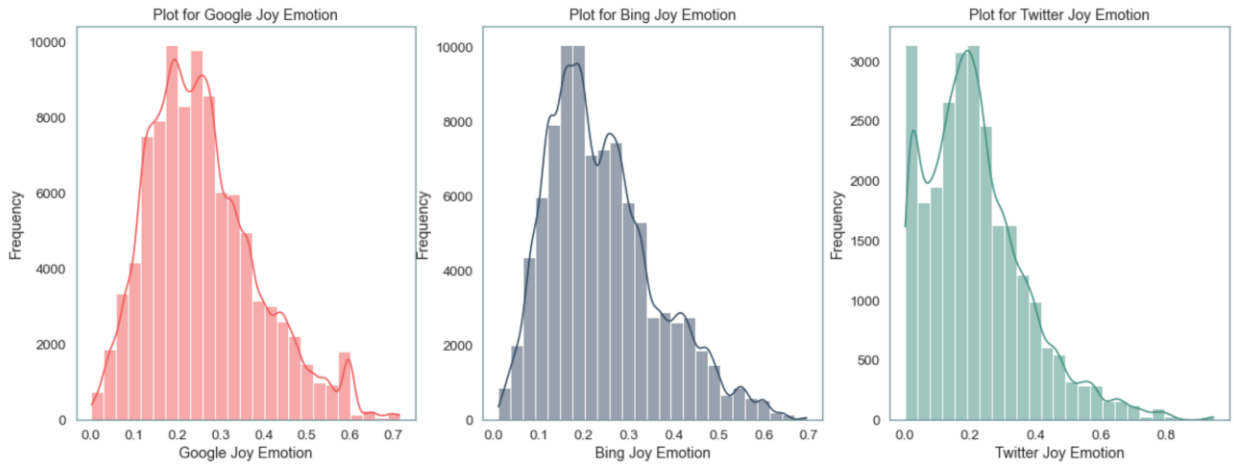


FIGURE 16: ***HISTOGRAM FOR EMOTION OF JOY ACROSS GOOGLE, BING, AND TWITTER***

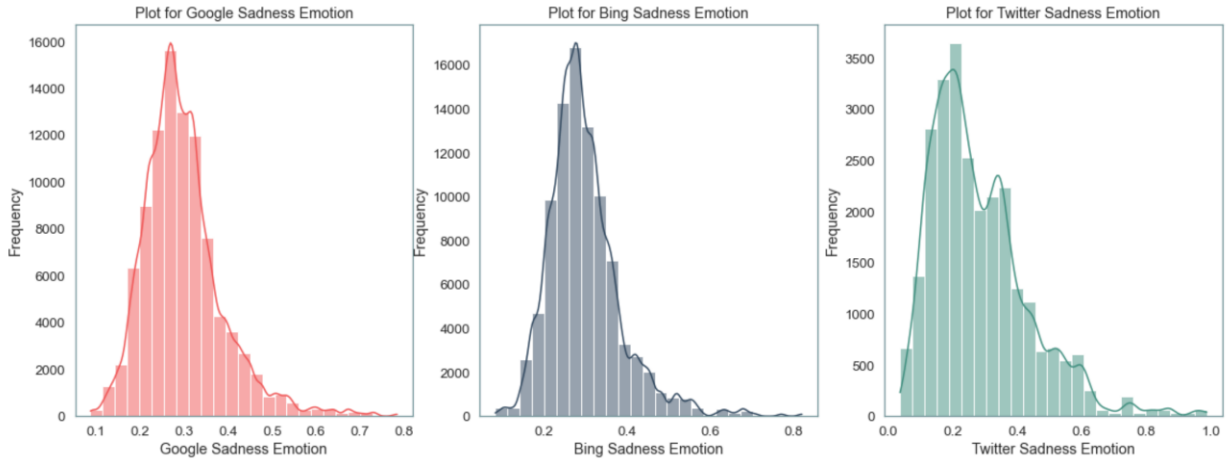


*FIGURE 17: **HISTOGRAM FOR EMOTION OF SADNESS ACROSS GOOGLE, BING, AND TWITTER***

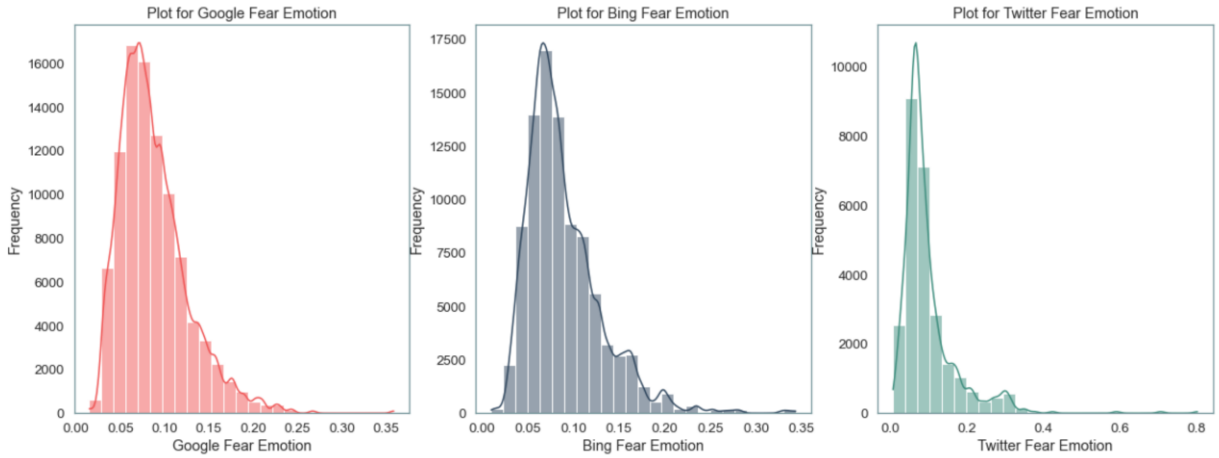


*FIGURE 18: **HISTOGRAM FOR EMOTION OF FEAR ACROSS GOOGLE, BING, AND TWITTER***

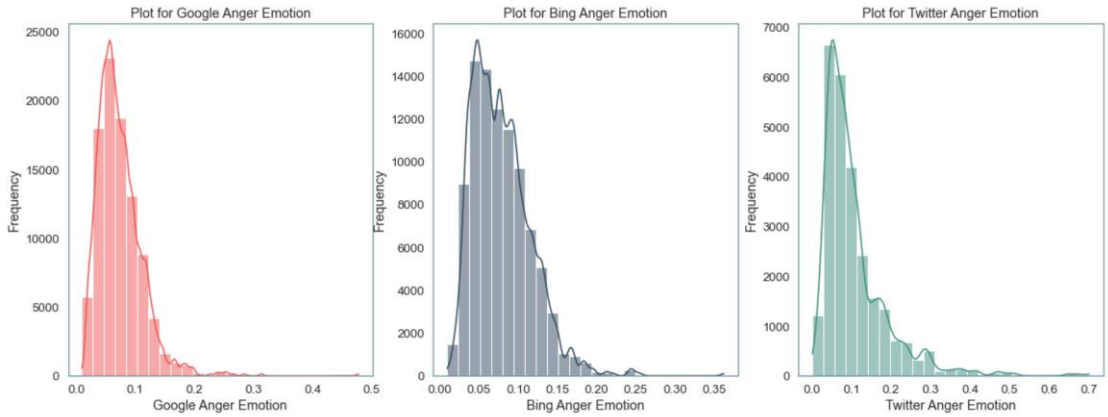


*FIGURE 19: **HISTOGRAM FOR EMOTION OF ANGER ACROSS GOOGLE, BING, AND TWITTER***

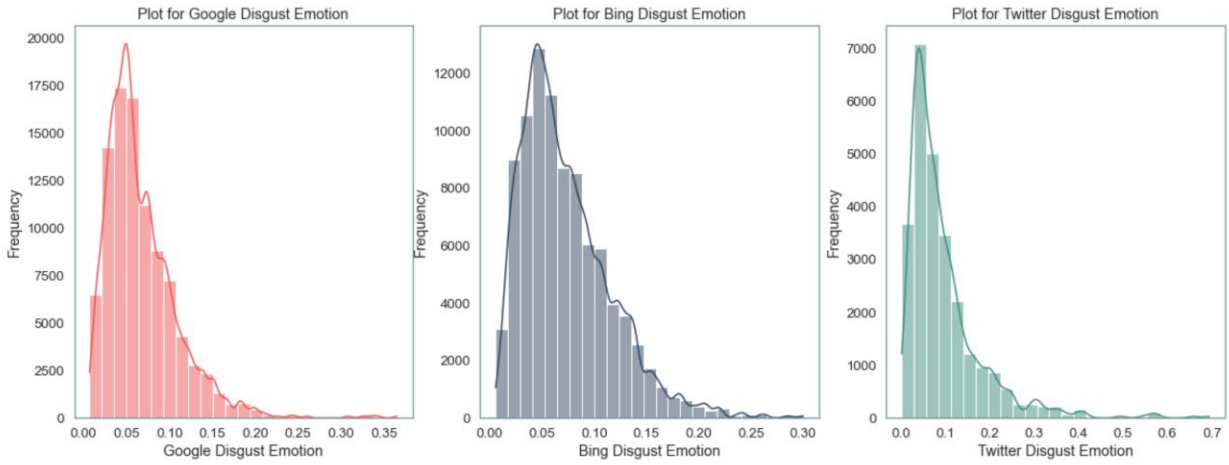


*FIGURE 20: **HISTOGRAM FOR EMOTION OF DISGUST ACROSS GOOGLE, BING, AND TWITTER***

TABLE 12: ***SENTIMENT OF SOCIAL AND SCIENTIFIC TOPICS ACROSS GOOGLE, BING, AND TWITTER.***

Sentiment on Social Topics

	Google	Bing	Twitter
Abortion	-0.475000	-0.499000	-0.469000
Anti-LGBT	-0.383000	-0.279000	-0.623000
Anti-Muslim violence	-0.454000	-0.502000	-0.459000
Extremism	-0.356000	-0.512000	-0.450000
Feminism	0.124000	-0.077000	-0.354000
Gay Marriage	0.060000	-0.132000	-0.440000
Gender Inequality	-0.003000	0.081000	-0.063000
Gun Control	-0.585000	-0.562000	-0.350000
Hate Crimes	-0.562000	-0.682000	-0.777000
Immigration	-0.134000	-0.271000	-0.464000
Online Censorship	-0.287000	-0.496000	-0.458000
Police Brutality	-0.326000	-0.369000	-0.271000
Privacy Rights	-0.004000	0.016000	-0.416000
Racism	-0.227000	-0.271000	-0.491000
Religious Freedom	-0.013000	0.014000	0.226000
Social Media	0.109000	-0.115000	-0.184000
Universal Healthcare	0.147000	-0.002000	-0.202000

Sentiment on Scientific Topics

	Google	Bing	Twitter
Artificial Intelligence	0.448000	0.510000	0.127000
Climate Change	0.022000	-0.096000	-0.275000
Coronavirus	-0.376000	-0.305000	-0.630000
Critical Race Theory	-0.233000	-0.264000	-0.243000
Depression	-0.428000	-0.379000	-0.580000
Marijuana	-0.114000	-0.308000	-0.336000
Suicide	-0.403000	-0.581000	-0.802000
Vaccines	-0.151000	-0.129000	-0.444000

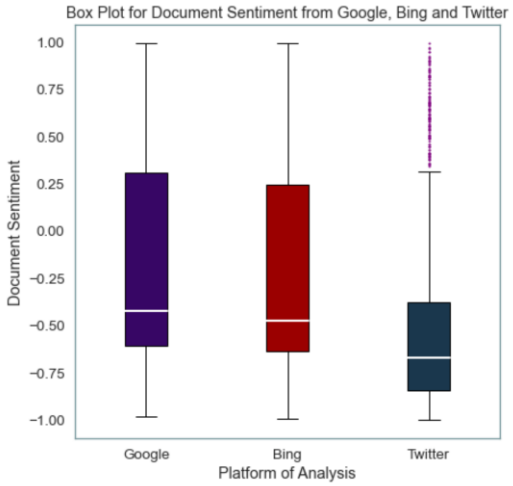


FIGURE 21: ***BOX PLOT SHOWING THE AVERAGE SENTIMENT, MEDIAN, AND THE QUARTILES FOR GOOGLE, BING, AND TWITTER.***

Joy Emotion on Social Topics

	Google	Bing	Twitter
Abortion	0.172000	0.152000	0.186000
Anti-LGBT	0.219000	0.235000	0.126000
Anti-Muslim violence	0.205000	0.187000	0.213000
Extremism	0.211000	0.180000	0.224000
Feminism	0.385000	0.325000	0.272000
Gay Marriage	0.281000	0.240000	0.223000
Gender Inequality	0.347000	0.371000	0.278000
Gun Control	0.166000	0.149000	0.154000
Hate Crimes	0.165000	0.133000	0.158000
Immigration	0.225000	0.191000	0.172000
Online Censorship	0.246000	0.219000	0.240000
Police Brutality	0.189000	0.189000	0.211000
Privacy Rights	0.305000	0.313000	0.292000
Racism	0.278000	0.247000	0.172000
Religious Freedom	0.276000	0.260000	0.312000
Social Media	0.375000	0.302000	0.204000
Universal Healthcare	0.303000	0.287000	0.230000

*TABLE 13: **KEYWORD WISE EMOTION OF JOY ACROSS THE THREE ISSUE TYPES.***

Joy Emotion on Scientific Topics

	Google	Bing	Twitter
Artificial Intelligence	0.418000	0.423000	0.451000
Climate Change	0.374000	0.348000	0.274000
Coronavirus	0.244000	0.248000	0.166000
Critical Race Theory	0.243000	0.252000	0.216000
Depression	0.232000	0.251000	0.206000
Marijuana	0.260000	0.220000	0.232000
Suicide	0.200000	0.156000	0.151000
Vaccines	0.271000	0.288000	0.273000

Joy Emotion on International Topics

	Google	Bing	Twitter
China	0.268000	0.255000	0.229000
Israeli-Palestinian Conflict	0.221000	0.216000	0.246000
Kashmir	0.283000	0.232000	0.287000
Taliban	0.223000	0.233000	0.210000
Russia-Ukraine	0.181000	0.168000	0.246000
Nuclear Weapons	0.211000	0.182000	0.119000

*TABLE 14: **KEYWORD-WISE EMOTION OF SADNESS ACROSS THE THREE-ISSUE TYPES.***

Sadness Emotion on Social Topics

	Google	Bing	Twitter
Abortion	0.305000	0.324000	0.330000
Anti-LGBT	0.306000	0.295000	0.351000
Anti-Muslim violence	0.278000	0.286000	0.287000
Extremism	0.293000	0.280000	0.286000
Feminism	0.243000	0.254000	0.195000
Gay Marriage	0.288000	0.263000	0.230000
Gender Inequality	0.286000	0.295000	0.309000
Gun Control	0.255000	0.269000	0.208000
Hate Crimes	0.269000	0.255000	0.189000
Immigration	0.300000	0.283000	0.245000
Online Censorship	0.279000	0.295000	0.275000
Police Brutality	0.248000	0.257000	0.176000
Privacy Rights	0.247000	0.244000	0.216000
Racism	0.304000	0.298000	0.299000
Religious Freedom	0.273000	0.258000	0.194000
Social Media	0.264000	0.263000	0.330000
Universal Healthcare	0.308000	0.297000	0.302000

Sadness Emotion on Scientific Topics

	Google	Bing	Twitter
Artificial Intelligence	0.245000	0.239000	0.190000
Climate Change	0.237000	0.246000	0.186000
Coronavirus	0.341000	0.328000	0.252000
Critical Race Theory	0.274000	0.265000	0.238000
Depression	0.486000	0.464000	0.576000
Marijuana	0.301000	0.285000	0.263000
Suicide	0.398000	0.406000	0.397000
Vaccines	0.339000	0.330000	0.324000

Sadness Emotion on International Topics

	Google	Bing	Twitter
China	0.294000	0.277000	0.292000
Israeli-Palestinian Conflict	0.268000	0.280000	0.242000
Kashmir	0.317000	0.311000	0.260000
Taliban	0.334000	0.323000	0.410000
Russia-Ukraine	0.320000	0.295000	0.278000
Nuclear Weapons	0.257000	0.269000	0.289000

Anger Emotion Table

	Google	Bing	Twitter
Social	0.081649	0.087954	0.108240
International	0.082798	0.088565	0.117268
Scientific	0.056251	0.060739	0.080983
Overall	0.075317	0.081049	0.102954

Joy Emotion Table

	Google	Bing	Twitter
Social	0.255723	0.234015	0.215706
International	0.231169	0.214406	0.222791
Scientific	0.280257	0.273203	0.246208
Overall	0.257302	0.240333	0.224948

Sadness Emotion Table

	Google	Bing	Twitter
Social	0.279099	0.277288	0.260134
International	0.298369	0.292433	0.295338
Scientific	0.327591	0.320413	0.303159
Overall	0.295343	0.291348	0.278051

Fear Emotion Table

	Google	Bing	Twitter
Social	0.076344	0.074255	0.083879
International	0.107307	0.113649	0.109958
Scientific	0.092419	0.095387	0.111939
Overall	0.086485	0.087333	0.096168

TABLE 15: ***THE FOUR EMOTIONS ACROSS GOOGLE, BING, AND TWITTER FOR THE THREE-ISSUE TYPES.***

TABLE 16: ***THE TOP 10 PUBLISHERS FOR GOOGLE AND BING, ALONG WITH THEIR FREQUENCIES.***

Top 10 Publishers

	Google	Frequency (Google)	Bing	Frequency (Bing)
0	The New York Times	6992	Fox News	3978
1	Reuters	3802	The New York Times	3912
2	Al Jazeera	3491	Associated Press	3671
3	Washington Post	3344	Reuters	2589
4	NPR	2872	The Guardian	2567
5	The Guardian	2580	CNN	1832
6	The Hill	2437	Forbes	1809
7	CNN	1973	ABC News	1781
8	ABC News	1964	Washington Post	1442
9	The Wire	1822	Washington Observer	1421

*TABLE 17: **THE TOP 10 PUBLISHERS FOR GOOGLE AND BING, ALONG WITH THEIR FREQUENCIES AND POLITICAL LEANINGS.***

Top 10 Publishers

	Google	Frequency (Google)	Ratings (Google)	Bing	Frequency (Bing)	Ratings (Bing)
0	The New York Times	6992	left-center	Fox News	3978	right-center
1	Reuters	3802	center	The New York Times	3912	left-center
2	Al Jazeera	3491	left-center	Associated Press	3671	center
3	Washington Post	3344	left-center	Reuters	2589	center
4	NPR	2872	center	The Guardian	2567	left-center
5	The Guardian	2580	left-center	CNN	1832	left-center
6	The Hill	2437	center	Forbes	1809	center
7	CNN	1973	left-center	ABC News	1781	left-center
8	ABC News	1964	left-center	Washington Post	1442	left-center
9	The Wire	1822	left	Washington Observer	1421	right-center

*FIGURE 22: **SCATTER PLOT SHOWING THE NUMBER OF PUBLISHERS LEANING TOWARDS PARTISANSHIP ACROSS GOOGLE AND BING.***

API Keys and Password

Part of the data is hosted in Github, but Github has a size limit on files, so it might have some data inconsistencies and unexpected bugs. The given SQL File will create the required tables and dump the data into it as well, making local running much simpler.

For the application, there are API Keys and Passwords that cannot be made public. So, I have included it here. The location to add these APIs into is mentioned in the Readme file.

For Azure Database in azure_config.py

```
azure_pwd = '{SearchBias@777}'
azure_id = 'batsalg'
azure_svname = 'capstone-search.database.windows.net'
```

For IBM Servers in ibm_config.py

```
url_api = 'aEgAo9_wTdLPk1gklVeKHukDzkXp5gXote0yttpMPrC0'
url_ibm = 'https://api.eu-gb.natural-language-
understanding.watson.cloud.ibm.com/instances/7fade06f-33fa-4aa6-8e0d-
64acca721fa4'
```

For Twitter API in twitter_config.py

```
consumer_key = "uJpYtY7l0NkhAlKEGI3KAtzdr"
consumer_secret = "fCXU3SKcRd0oS0QOTl7we35ibzC2NfWlbW1WJa13W3iefJkEEW"
access_token = "883611374697332737-usBDXrTpU8Wic95w612p2H74sAoC88Y"
access_token_secret = "HIwSQmYLoQKMBABkZDxsWHR85P0PAsNsIlVxPSQZ6S7Lu"
```

THE END