STEM Fellowship Big Data Challenge 2021

The Plebeian Algorithm

A Democratic Approach to Censorship and Moderation

Benjamin D. Fedoruk[1], Harrison S. Nelson[2], Kai A. Fucile Ladouceur[3], and Russell M. Frost[4]

[1]University of Ontario Institute of Technology
[2]Queen's University
[3]Confederation College
[4]Lakehead University

November 18, 2021

Abstract

The infodemic created by the COVID-19 pandemic has created several societal issues, including a rise in distrust between the public and health experts and even a refusal of some to accept vaccination; some sources suggest that 1 in 4 Americans will refuse the vaccine.[1] This social concern can be traced to the level of digitization today – particularly in the form of social media. The goal of the research was to determine an optimal social media algorithm, one which is able to reduce the number of cases of misinformation, and which also ensures that certain individual freedoms (such as the freedom of expression) are maintained. After performing the analysis described herein, an algorithm was abstracted. The discovery of the set of abstract aspects of an optimal social media algorithm was the purpose of the study. As social media was the most significant contributing factor to the spread of misinformation, the team decided to examine infodemiology across various text-based platforms (Twitter, 4chan, Reddit, Parler, Facebook, and YouTube). This was done by utilizing a sentiment analysis to compare general posts with key terms flagged as misinformation (all of which concern COVID-19) to determine their verity. In gathering the datasets, both APIs (installed using Python's 'pip' tool) and also pre-existing data compiled by standard scientific third parties were used. The sentiment can be described using bimodal distributions for each platform, with a positive and negative peak, as well as a skewness. It was found that in some cases, misinforming posts can have up to 92.5% more negative sentiment skew compared to accurate posts. From this, the novel Plebeian Algorithm is proposed, which utilizes sentiment analysis and post popularity as metrics to flag a post as misinformation. This algorithm diverges from that of the status quo, as the Plebeian Algorithm uses a democratic process to detect and remove misinformation. A method was constructed in which content deemed misinformation to be removed from the platform is determined by a randomly selected

jury of anonymous users. This not only prevents these types of infodemics, but also guarantees a more democratic way of using social media that is beneficial for repairing social trust and encouraging the public's evidence-informed decision making.

Keywords: infodemiology, misinformation, algorithm, social media, plebeian

1 Introduction

The internet is a powerful tool for spreading information; as such, it follows that it is equally powerful for spreading misinformation. In 2019 the number of social media users worldwide was 3.484 billion,[2] with that number increasing year-by-year, by an average of 9%.[2,3] With this increased usage, the "power-user" or microinfluencer phenomenon has arisen, where popular social media accounts are able to reach large numbers of readers. This is increasingly important as more people begin to use social media as a source for news.[4] This news is third-party by popular influencers, not posted or moderated by the social media companies themselves. Past analyses examining online misinformation often classify posts as misinformation using a "Point-And-Shoot" algorithm: this is the status quo. However, some algorithms will be better at combating misinformation than others. The Plebeian Algorithm creates a criteria that social media websites should take into account when designing their algorithms in order to reduce misinformation. This reduction of misinformation is thought to be achieved by examining the correlation between sentiment and misinformation; it has been found that posts containing misinformation tend to be of more negative sentiment when compared categorically to other posts covering the same issue.[5] Due to this correlation, it is hypothesized that an algorithm which encourages positive interactions will also reduce the amount of misinformation present on the platform through a democratic manner.

Misinformation is a key problem, yet many terms are confused in studies. Herein, the authors shall define several key terms which are often used interchangeably, but whose definitions are specific and distinct. Firstly, misinformation shall be defined as the spread, intentional or otherwise, of false information.[6] The intentions of the individual(s) spreading the information is irrelevant. Secondly, disinformation is the purposeful spread of false information.[6] A similar yet distinct definition is malinformation, which is the malicious spread of false or misleading information.[6] Finally, fake news is defined as any misinformation (with or without intention), which readers interpret as trustworthy news.[6] For the purpose of this study, misinformation will be studied in depth. Infodemics can additionally be applied to the realm of health care; infodemics have the potential to intensify outbreaks when there is uncertainty among the public concerning evidence-informed preventative and protective health measures.[7]

Prior to investigating the spread of misinformation, it is pertinent to define the concept of infodemiology and misinformation. This research paper defines an infodemic according to the World Health Organization (WHO) as "too much information including false or misleading information in digital and physical environments during a disease outbreak," which "causes confusion and risk-taking behaviours that can harm health [and] also leads to mistrust in health authorities and undermines the public health response,".[7] The study of the spread of infodemics on a large scale,

especially pertaining to medical misinformation, is known as infodemiology. Infodemics have the potential to intensify outbreaks when there is uncertainty among the public concerning evidence-informed preventative and protective health measures".[7] The WHO has additionally linked the rapid surge of such infodemics during the COVID-19 pandemic to "growing digitization" which can support the global reach of information but can also quickly amplify malicious or fabricated messages.[7] The second relevant definition is that of the concept of misinformation, which has been defined similarly to the definition used by the WHO in relation to the infodemic,[7] but specifically refers to the distinct lack of verity in information related to a specific field.

At their core, most social media websites aim to maximize the amount of time that users spend on their platforms. This maximization of user page-time leads to companies utilizing highly specialized and trained machine learning to advertise content on users' feeds.[8] At the same time, this can have unintended adverse effects such as maximizing the time a user engages with content that is not verified for accuracy. The proposed solution to this disparity between engagement and integrity is to create democratically moderated spaces. Democratic spaces and recommendations to posts with more positive sentiment are integral concepts in the Plebeian Algorithm, based on the latest evidence that misinformation tends to be more negative.[5] The Plebeian Algorithm is an algorithm described herein, for the purpose of the control of the spread of misinformation on social media. It is extremely beneficial compared to other existing algorithms. The currently implemented point-and-shoot algorithms are hyper-tuned to specific sources of misinformation, surrounding specific topics. However, they are not adaptable to the fluidity of the definition of true information.

As mentioned earlier most social media platforms work on a model similar to Twitter, Facebook, or YouTube where content is recommended based on user engagement;[8] however, this is not true to the same extent for all websites. One example of a website that breaks the expectations for social media algorithms is 4chan. 4chan is an excellent epitome of ephemeral social media, where content is completely anonymous and is rapidly discarded regardless of popularity;[9] in addition, there is close to no moderation and the content tends to be more negative in sentiment. This is also exemplified in Parler, an alternative social media platform established in September of 2018, that aimed to bring forth a platform with total freedom of speech. As a consequence, Parler attracted those who were banned from other social media websites creating "echo chambers, harbouring dangerous conspiracies and violent extremist groups",[10] such as those who were involved in raiding the U.S. Capitol on Jan. 6th, 2021. Reddit also has a forum-based system similar to 4chan. However, individual fora on these platforms have moderators who work to combat negative sentiment throughout the website. Reddit's issue lies in its incredibly isolated fora, as tailoring one's feed to being a vast majority of explicitly handpicked fora is a part of the experience; this allows for some fora to have little to no moderation.[11]

There have been several related works of research in the field of the detection of misinformation on social media platforms. These works include the studies of the connection between misinformation and cognitive psychology,[12] the analysis of geospatial infodemiology,[13] the effect of recommendation algorithms on infodemiology,[14] the use of distributed consensus algorithms to curb the spread of misinformation,[15] and the naming conventions used for viruses.[16] Although these works are in

alignment with this study, they do not propose the same solution. The study, which offers a solution closest to that proposed by the Plebeian Algorithm, discusses the efficacy of curbing the spread of misinformation through layperson judgements.[17] Notably, this work discusses the merits surrounding a layperson algorithm, but does not make suggestions for its implementation.

The objective of the study will be to determine the optimal social media algorithm to reduce the spread of misinformation, while ensuring personal freedoms. The investigation conducted in this paper will have far-reaching implications which will alter how misinformation in social media platforms is addressed.

The three major implications include:

- The creation of a more open and democratic environment on social media platforms;
- An overall reduction in political divisiveness and extremist sentiment both on- and off-line; and,
- An increase in informed users who can make well-informed opinions on subjects.

2 Materials & Methods

A detailed step-based methodology was used to analyze data throughout the research process. Python 3.9 was the language of choice through all aspects of the project. All libraries used can be accessed using `pip`. The visualization of data was performed using the `matplotlib` and `seaborn` libraries in Python. APIs were used from Twitter, Reddit, and 4chan, gathering data regarding username, date, post and text. Furthermore, two datasets were gathered from academic sources, containing post data from Twitter[18] and Parler.[10] Various Python libraries were used to interact and connect with the APIs, including: `twarc`, `urllib3`, and `basc_py4chan`. The following Python libraries were used to clean the data: `beautifulsoup4`, `demoji`, and `pyenchant`. The `pandas` library for Python was used to retrieve and store third-party datasets,[10,18–22] and the `numpy` library was used for various array operations. Finally, the `nltk` library was used to perform sentiment analysis, and `sklearn` was used to perform regressions.

Python was selected due to its ease of connectivity to the various APIs; it is well-supported among a strong community, and as such, connecting to various APIs was done through pre-written libraries. This reduced the programming time, while increasing the efficiency and reliability of the code.

In three of the social media services for which APIs were used (i.e. Twitter, Reddit and 4chan), four steps were performed: (1) gather data using the API and the associated Python library; (2) clean data to create a Python set of strings, containing no URLs (removed using regular expressions), HTML (removed using `beautifulsoup4`), usernames (removed using regular expressions), emojis (replaced with text using `demoji`), or non-English language (removed using `pyenchant`); (3) perform sentiment analysis using `nltk`'s `SentimentIntensityAnalyzer` class; and, (4) save the cleaned and sentiment analyzed data frame as a pickle file. Then, a visualization script was programmed to display the sentiment data gathered from the social media post data. To ensure confidentiality of users, only aggregate data was

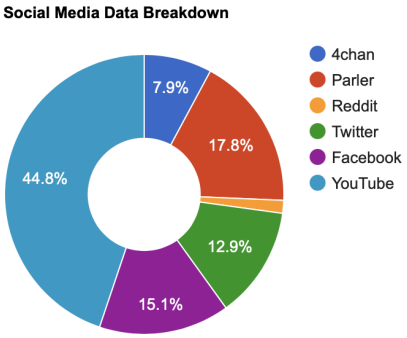

Figure 1: Social Media Data Breakdown

displayed. Plotting a histogram with a KDE resulted in the various graphs produced by the research team. Data for which sentiment analysis returned inconclusive due to textual limitations was removed from visualization. Limiting the language to English has the benefit of statistical comparison congruence. One notable platform which did not use an API is Facebook. The reason for this is due to the restrictions placed on the Facebook API, in terms of depth and breadth of research.

The six social media services analyzed (4chan, Twitter, Parler, Reddit, YouTube and Facebook) had various amounts of associated data. The data breakdown is described in 1.

The sentiment analysis dictionary selected for the analysis performed herein is the Valence Aware Dictionary and sEntiment Reasoner (VADER). This dictionary was selected as it is the industry standard for a wide array of general statement analyses, and especially recognized for producing highly accurate results with social media platforms. As such, VADER is the optimal dictionary for the purposes of this research. Although ideal algorithms should implement various checks and balances for the sentiment analysis system implemented, this paper shall focus on strictly the VADER dictionary, which is solely positive and negative sentiment. Other sentiment analysis tools exist to examine specific emotions (including anger, fear, surprise, happiness, etc.).

Keyterm analysis was used for the data cleaning process, in order to determine which strings were classified as relating to a specific topic. The keyterms were gathered using a list of the most commonly held terms, gathered from Twitter, which were directly associated with misinformation.

In order to confirm the academic literature[5] regarding the correlation between negative sentiment and verity of information, analysis was performed using Twitter. Data was filtered, such that only Tweets containing a set of potentially misinformative keywords were assigned to be assessed using a sentiment analysis. Both were plotted through histogram, and the KDEs were compare (relative to each respective maxima).

Misinformation is directly correlated to negativity. A misinformative post is often negative in sentiment. However, this is not a certainty. As such, when determining an optimal algorithm, it will be critical to use sentiment analysis to narrow the potential misinformative candidates, and then to use further methodology – a jury process – to accurately detect misinformation.

The study defines several mathematical terms. Many of the histograms and KDEs as described above form a bimodal distribution. The polarity score upon which the two peaks are centered is termed μ^+ and μ^-, where the sign indicates whether the term refers to the positive or negative peak. The other variable defined is the skewness of the distribution as a whole, which is described using the symbol γ. When the positive peak is the major mode, then $\gamma \in (0, \infty)$. Contrarily, when the positive peak is the minor mode, then $\gamma \in (-\infty, 0)$. The frequency function f describes the frequency curve represented by the KDE (such that $f(p)$ represents the frequency of strings with polarity score p). The skewness is calculated using the following equation:

$$\gamma = \frac{2f(\mu^+) - 2f(\mu^-)}{\mu^+ f(\mu^+) + \mu^+ f(\mu^-) - \mu^- f(\mu^+) - \mu^- f(\mu^-)} \tag{1}$$

This equation for skewness was derived using the following derivation:

$$\gamma = \frac{\Delta f(\mu)}{\Delta \mu} \div (f(\mu))_{av}$$
$$\gamma = \frac{\Delta f(\mu)}{\Delta \mu} \div \frac{\Sigma f(\mu)}{2}$$
$$\gamma = \frac{2\Delta f(\mu)}{(\Delta \mu)(\Sigma f(\mu))}$$
$$\gamma = \frac{2(f(\mu^+) - f(\mu^-))}{(\mu^+ - \mu^-)(f(\mu^+) + f(\mu^-)}$$
$$\gamma = \frac{2f(\mu^+) - 2f(\mu^-)}{\mu^+ f(\mu^+) + \mu^+ f(\mu^-) - \mu^- f(\mu^+) - \mu^- f(\mu^-)}$$

As will be described throughout the progression of the research, there is a strong connection between the sentiment of a post, and the

3 Results

The results of the analysis will be divided by the social media platform. They will be presented in the following order:

1. Reddit
2. 4chan
3. Facebook
4. YouTube
5. Parler
6. Twitter

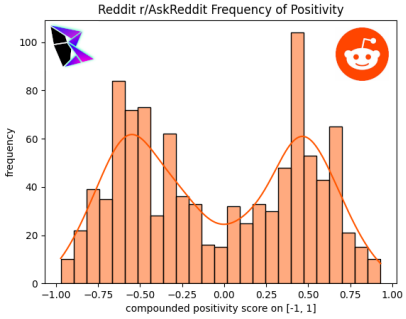

Figure 2: Reddit r/AskReddit Frequency of Positivity Histogram

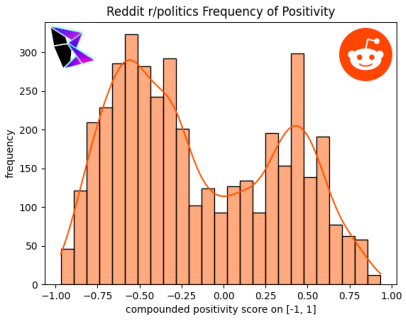

Figure 3: Reddit r/politics Frequency of Positivity Histogram

3.1 Reddit

When analyzing Reddit's data, a series of Subreddits were selected. The Subreddits selected were: r/AskReddit, r/AskThe_Donald, r/conspiracy, r/covid, r/kindness, r/movies, r/politics, and r/EnoughTrumpSpam. These subreddits were selected as an array of options, allowing an analysis of probable misinformative, probable truthful, and unknwon sources. The data was gathered using the Python library `urllib3`. The first Subreddit to be examined herein is r/AskReddit. This Subreddit tends to contain a wide variety of posts from a myriad of conversation topics. As such, it is relatively indicative of Reddit on the whole. r/AskReddit's histogram can be found in Figure 2. The bimodal distribution has $\mu^- \approx -0.54$, $\mu^+ \approx 0.48$, and $\gamma \approx -0.03214$.

It should be noted that the extremal frequencies of the bimodal distribution are approximately equal between the negative and positive peaks. Another notable Subreddit examined was r/politics, which provided a sample of posts potentially swayed by political leaning of the Reddit users. The histogram and KDE for this analysis is displayed in Figure 3. The bimodal distribution for r/politics has $\mu^- \approx -0.56$, $\mu^+ \approx 0.43$, and $\gamma \approx -0.37776$.

r/politics' contents have a stronger negative skew, as is apparent by the KDE. The final Subreddit to be examined is an avant-garde Subreddit: r/conspiracy. In this community, users share various conspiracy theories. When one scrolls through r/conspiracy, plenty of misinformation can easily be noted, including misinformation surrounding Flat Earth Theory and QAnon. The histogram for r/conspiracy is found in Figure 4. The bimodal distribution for r/conspiracy has $\mu^- \approx -0.56$, $\mu^+ \approx 0.39$, and $\gamma \approx -0.33904$.

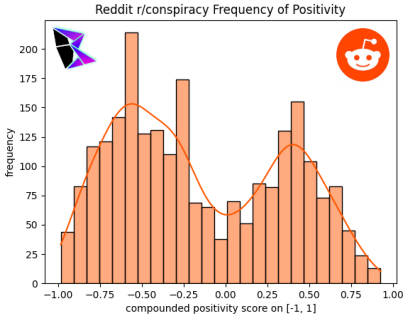

Figure 4: Reddit r/conspiracy Frequency of Positivity Histogram

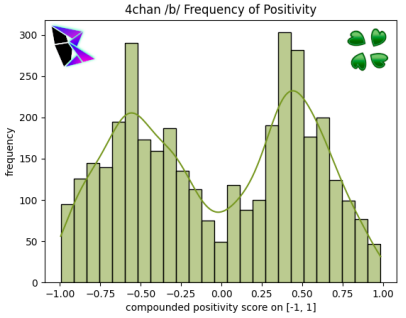

Figure 5: 4chan /b/ Frequency of Positivity Histogram

As can be noted by r/conspiracy, the conspiratorial posts (which are known to contain a large volume of misinformation) are more often negative. This can be noted due to the difference in the peaks of the bimodal distribution.

3.2 4chan

In order to analyze the 4chan data, five boards were selected: /b/, /a/, /v/, /pol/, and /r9k/. These five boards were selected due to their highest post frequency compared to other 4chan boards. The data was gathered using `basc_py4chan`. For each of these boards, a histogram was plotted (with an overlayed KDE) with 30 bins. A visualization for the histogram for /b/'s sentiment can be found in Figure 5. /b/ is described as the Random board, containing a wide mixture of conversation from across 4chan. The bimodal distribution for /b/ has $\mu^- \approx -0.55$, $\mu^+ \approx 0.46$, and $\gamma \approx 0.11380$.

Another board to be visualized in this report is the visualization for the sentiment of /pol/, which can be found in Figure 6; /pol/ contains political discussion. The bimodal distribution for /pol/ has $\mu^- \approx -0.61$, $\mu^+ \approx 0.38$, and $\gamma \approx -0.16559$.

It should be noted that the levels of extreme negative sentiment (i.e. with a polarization score of less than 0.75) are dramatically higher in /pol/ compared to /b/. This demonstrates that political topics tend to be more negative on 4chan.

Overall, it should be noted that 4chan consistently contains a large number of negative posts, which is greatly dependant upon the topic of the board. Boards which pertain to specific recreational activities (such as /v/ for video games or /a/ for anime) have a lesser degree of negative polarity.

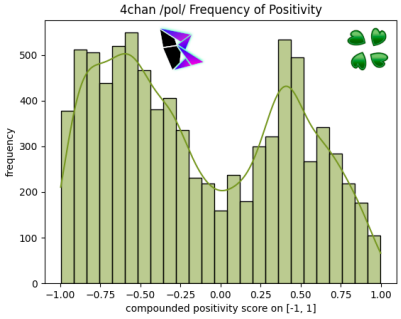

Figure 6: 4chan /pol/ Frequency of Positivity Histogram

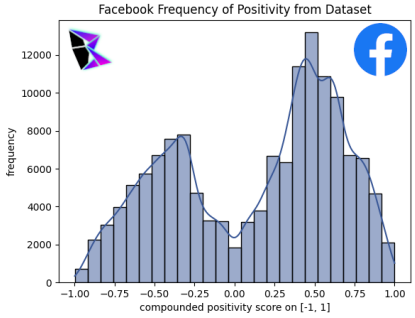

Figure 7: Facebook Frequency of Positivity Histogram

3.3 Facebook

It is pertinent for this paper to perform analysis on Facebook to perform analysis on Facebook, which is currently the social platform with the largest user base of 2.8 billion active monthly users.[23] Facebook has proven to be the social media platform with the highest user base, and as such it is pertinent for this paper to perform analysis on data collected for Facebook. A dataset that specifically contains data predating the COVID-19 pandemic was accessed to broaden the scope of the sentiment analysis.[24] The dataset includes data gathered from Facebook's inception until 2017. A dataset with a random selection of Facebook comments from the temporal range[24] was selected for sentiment analysis using VADER.

In Figure 7, a histogram was plotted with 30 bins, depicting the frequency of Facebook comments at various sentiment analytic levels. A KDE was overlayed onto the plot to show the general trend.

Notable features of the bimodal histogram include the sharp positive peak and wide negative peak. It should be noted for the integral for the KDE is as follows:

$$\int_{-1}^{0} p(x)dx \approx \int_{0}^{1} p(x)dx.$$

Furthermore, the following values have been extracted from the KDE: $\mu^{+} \approx 0.43$, $\mu^{-} \approx -0.29$ and $\gamma \approx 0.49858$.

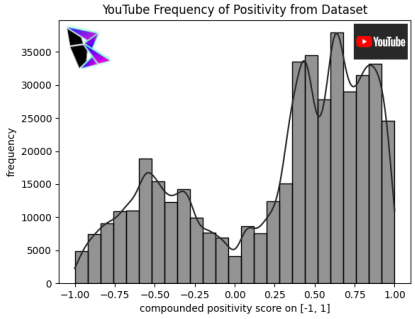

Figure 8: YouTube COVID-19 Frequency of Positivity Histogram

3.4 YouTube

YouTube is based entirely on long-form video content and tends itself towards more in-depth topics. A pre-selected dataset of YouTube comments,[25] after sentiment analysis, has been visualized and presented in Figure 8.

The dataset was collected in 2017, and as such does not contain misinformation related to COVID-19. This helps to broaden the temporal scope of this analysis and ensure that the trends present hold in data outside of the COVID-19 pandemic (i.e. prior to January 2020). It was also limited in geographic scope to the United States, the United Kingdom and Canada. This limitation was due to the availability of data. It should be noted that these three countries represent the English-speaking members of the Group of Seven (G7), a group of the seven most democratic, affluent and pluralist nations in the world.

As can be noted, there is a strong positive skewness in the data, with $\mu^+ \approx 0.67$, $\mu^- \approx -0.52$, and $\gamma \approx 0.66593$. The high positive skewness should be noted for these YouTube comments. Potential explanations for this trend will be discussed in a later section.

3.5 Parler

An analysis of Parler is a transition from the traditional analyses of Reddit and 4chan, due to the fact that Parler is not broken down into communities to which users subscribe, but is a single newsfeed-style system. The analysis of Parler should be contrasted to the analysis of Twitter in the subsequent section, as users migrated from Twitter to Parler due to a perception of limitations placed on their freedom of expression on Twitter.

When analyzing Parler, data was collected into a dataset throughout the COVID-19 pandemic and the period of time surrounding the events of January 6th, 2021.[10] Figure 9 contains a visualization of the COVID-19-related parleys posted between January 2020, and March 2020. The bimodal distribution for /pol/ has $\mu^- \approx -0.53$, $\mu^+ \approx 0.45$, and $\gamma \approx 0.22063$.

3.6 Twitter

The majority of the analysis performed through this paper was on the social media service, Twitter. The reason for this is due to the high amount of data regarding misinformation on the platform, as well as the overall popularity of the platform as

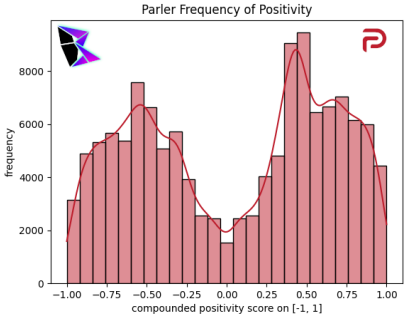

Figure 9: Parler COVID-19 Frequency of Positivity Histogram

a general case study, and the generality of the platform (compared to some other unorthodox data sources such as YouTube comments).

Similarly to Parler, Twitter Tweets are made at-large to the general public. There are no channels, boards, or Subreddits of any sort. However, due to the Twitter algorithm, there is an allowance of individuals' feeds to be in an echo chamber. Evidently, echo chambers should be avoided wherever possible. Echo chambers are a large contributor to the rampant spread of misinformation that is seen surrounding the COVID-19 pandemic.[26]

This study used a combination of both data gathered from the Twitter API and a dataset of pre-gathered COVID-19 Tweets.[18,19,21,22] The interface used to connect the Python code (and sentiment analysis) to the Twitter API was `twarc`. The reason for this duplication of analysis was to ensure that the data used was accurate. Precision must be maintained in both data collected by APIs and over a long duration.

In both of the studies (using the API and the dataset[18]), the study analyzed the broad sentiment of COVID-19-related Tweets, as well as filtered the data by keyword. The keywords used included terms concerning misinformation surrounding COVID-19, including "China Virus", "Bioweapon", and "Microchip". The filtered data then underwent sentiment analysis. Both of the sentimentally analyzed data were plotted on the standard histogram with overlaid KDE.

For the discussion, the study focused on the data gathered from the Twitter API, since a similar methodology was used for gathering data for the other social media platforms studied. However it should be noted that similar results are attained using the dataset.[18] Below in Figure 10 is a graph of the Twitter APIs gathered Tweets pertaining to COVID-19 (the broad topic), where the sentiments of the Tweets are plotted on a histogram with a KDE. A random sample of the data was taken for this analysis, as there were too many Tweets to reasonably analyze the population. The bimodal distribution has $\mu^- \approx -0.36$, $\mu^+ \approx 0.47$, and $\gamma \approx 0.86500$.

As can be noted, the positive peak for the KDE is nearly double the negative peak. This indicates that the number of positive Tweets far exceeds the number of negative Tweets. Comparatively in Figure 11, the negative peak of the bimodal distribution is on-par with the positive peak. This figure is a histographic representation of the polarity score for Tweets, after being filtered. The Tweets selected only contain terms which are known to be pertaining to COVID-19 misinformation. The bimodal distribution has $\mu^- \approx -0.42$, $\mu^+ \approx 0.47$, and $\gamma \approx 0.06420$. Thus, $\Delta\gamma \approx 0.80080$ between the two Twitter measurements.

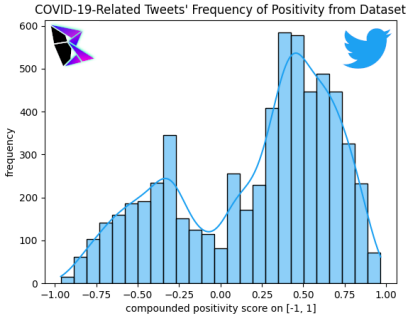

Figure 10: Unfiltered COVID-19 Twitter API Frequency of Positivity Histogram

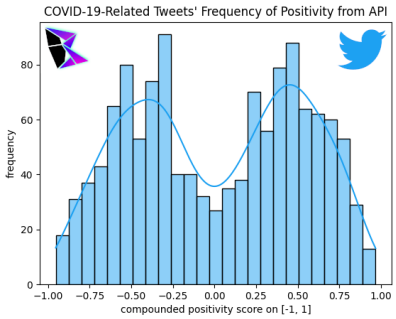

Figure 11: Filtered COVID-19 Twitter API Frequency of Positivity Histogram

The study's discussion of the reasoning behind this proportional increase in negative peak compared to positive peak will be discussed further in the subsequent section. Again, it must be noted that the same results can be seen when performing the identical analysis on the data gathered from the dataset.[18]

4 Discussion

In the discussion, not only will an analysis of the results and errors be explored, but the the Plebeian Algorithm and its benefits are also discussed as well as how it compares to the algorithms of the social media platforms studied.

4.1 Results Analysis

Several critical notes must be made with regards to the analysis of the quantitative features produced in the Results section.

First, it is critical to note that 4chan was the only social media platform studied which had an overall positive γ, notably on the all-encompassing board of /b/. It was gathered by the observations that a system which provided users with the freedom to determine which content got promoted – as opposed to an artificial intelligence algorithm – improved the sentiment of the average post. This is a key point in the Plebeian Algorithm, which is described in a subsequent section. Second, Twitter had a more moderate skew (i.e. closer to 0, or neutral) μ^-, indicating that users tended to be more positive than users on the other social media platforms analyzed in the study.

It is also critical to recall that there was a strong correlation between polarity scores as determined by a sentiment analysis algorithm, and the verity of the information communicated.[5] As such, the analysis provided herein can be applied to both the sentiment and the verity of a social media post.

4.1.1 Echo Chambers

In analysing the data represented by the KDEs and the skew of the bimodal distribution of the data towards negative sentiment, it is clear to see a confirmation of the Echo Chamber Thesis (a theory which states that "the Internet has produced sets of isolated ideologically homogeneous echo chambers, where similar opinions reinforce each other and lead to attitude polarization"[27]) as negative sentiment has a clear association with emotions such as anger, which have been shown to "...[reinforce] echo chamber dynamics [...] in the digital public sphere".[27] In fact, other studies have also predicted the link to this effect to be the impact of the specific algorithms used in the virtual space.[28] The link gives strong evidence to suggest that the algorithms currently being used by social media companies are creating the optimal medium through which misinformed opinions and content can grow and go uncontested. These echo chambers ensure that users are unable to get access to arguments which conflict with their beliefs, and expand their perspectives.

4.2 Sources of Error

Although the study attempted to limit error, there remained several sources of error stemming from the methodology of analysis used. The first source of error is the trouble with using key term searching, as it would not only give us results of posts of individuals spreading misinformation but also those trying to bring attention to issue of misinformation and those who spread it. Furthermore, the sentiment analysis would also have been unable to differentiate between a misinformed post and one that tried to bring attention to the problem. This is due to the fact that some of the true Tweets demonstrate overall negative emotions. The final problem with the key term search is that some of the key terms determined to be misinformative may in future be proven to be accurate information.

A further source of potential error comes in the form of the social media platforms utilised. One problem is that only four social media platforms were assessed, thus limiting the scope of the study. It very well could be that the trends found in the research will not show up in other social media platforms (e.g. Facebook). Another issue from the sources assessed was that they were all only textually based and thus the method proposed may not be replicable for more graphically based social media platforms (e.g. YouTube, Instagram, Snapchat).

4.3 Definition and Implementation of the Plebeian Algorithm

As described previously, the Plebeian Algorithm is a novel algorithm for identifying and removing misinformation through democratic means. It works in two distinct phases: the Flag Phase, to determine which posts are misinformative; and the Jury Phase, to judge the information to determine if removal is appropriate.

4.3.1 Flag Phase

The Flag Phase is tasked with the determination of possible misinformed posts. In doing so, the algorithm selects posts which have a large number of views, and then performs sentiment analysis on both the original post, as well as a "without-replacement simple random sample"[29] of comments or replies to the post. If the overall sentiment leans negative, then the post is flagged as being potentially misleading. The posts flagged will then be passed into Jury Phase.

4.3.2 Jury Phase

The Jury Phase is tasked with the trial and removal of truly misinformative posts. These posts are removed from the homepage or newsfeed of the user. During the Jury Phase, flagged posts are sent to a random selection of anonymous users known as jurors. This selection should provide a diverse group, consisting of varying political opinions to give the post a fair trial. The selection of jurors uses a "without-replacement simple random sample" of a population.[29] The number of jurors selected is exactly 10% of the number of viewers of a post, rounded up. It should be noted however that jurors are not forced to participate or vote. It is assumed that the number of voting jurors will be far less than the total number of jurors. Thus, selecting 10% of the population allows room for the uncertainty of juror engagement. The jurors are then asked to vote either for or against the removal of the post. Once the deliberation has lasted a set duration (or a threshold of response has been met), the results will be counted and the post will remain on the site, or be removed by the algorithm.

For reference, a summative flowchart detailing the Plebeian Algorithm as described can be found in Figure 12. The areas coloured in magenta constitute the Flag Phase, while the areas coloured in mint green constitute the Jury Phase.

4.4 Existing Algorithms

In the following section, each of the algorithms will be detailed for Reddit, 4chan, Parler, and Twitter. These analyses will be based on academic journal articles.[9–11,30,31] It is pertinent to analyze these on a case-by-case basis, as it must be ensured that the user base remains loyal to the brand and platform.[32] Prevalent existing algorithms include the PageRank and Hits algorithms.[33]

Before dissecting the individual social media algorithms that are currently being used for the various platforms, it is critical to mention that most of these algorithms have an identical goal: to get users to stay on the platform, thus ensuring a continued revenue for their organization. This objective contradicts the goal of preventing misinformation from spreading on the platform, as preventing misinformation means censorship, resulting in a reduction of revenue. This, however, is not to indicate that the Plebeian Algorithm is of little value to social media entrepreneurs. It is critical to give note that most social media companies (with the exception of Parler) have already shown themselves interested in curbing and moderating their own social media platforms through their implementation of "Point-and-Shoot" algorithms as well as censorship of high-profile posts and accounts (e.g. Twitter banning @realDonaldTrump). However, as has already been described, these algorithms are not effective at accomplishing their mission of reducing misinformation, and further,

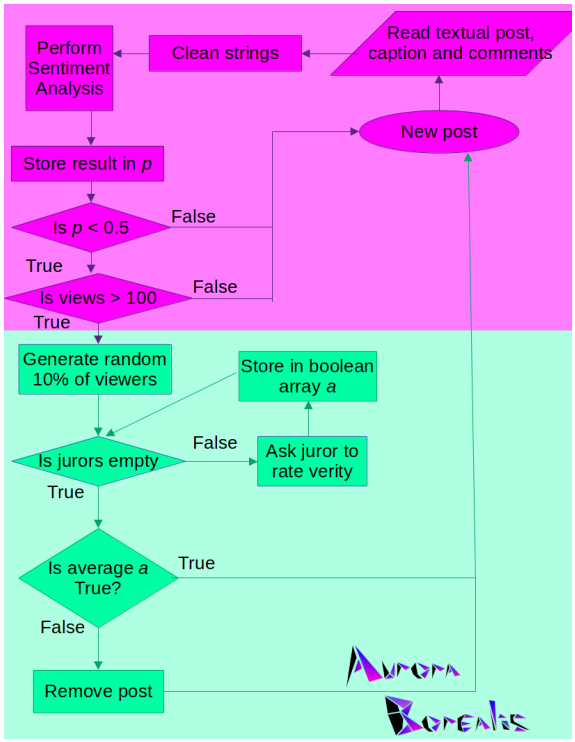

Figure 12: Flowchart of the Plebeian Algorithm

have caused users to become disillusioned with the service. This has led to many users joining platforms which capitalize on this disillusionment (e.g. Parler). Hence, by implementing the Plebeian Algorithm, these social media companies finally have a method which carefully balances moderation with freedom of expression that will re-inspire a sense of awe within their user base and bring back the notion of social media being a fun online space where people can collaborate and share freely.

4.4.1 Reddit

The algorithm used for Reddit is a simple upvote/downvote system, as was described in the Introduction. Users of Reddit are encouraged to upvote content they like and are encouraged to downvote content which they do not like. Posts with more upvotes are more widely shared whereas the opposite is true with posts with more downvotes. In Reddit, users are allowed to vote on both the original post, as well as any comments. "Comment trees" are inherently created by the system as users comment on comments (thereby chaining comments together into a tree-like formation).

The Reddit algorithm is tailored to the interests of the Reddit user. Through a system of subscriptions to various topics of conversations or Subreddits. Users will receive a mixture of content from the Subreddits to which they have subscribed, with additional, sporadic advertisement.

The system is essentially tailored to the specific user. This is in contrast to the Plebeian Algorithm, which emphasizes the democratic process for the determination of verity by the user base as a whole. Currently, Reddit contains no user-controlled means to fight misinformation aside from the "Report" button, which brings the issue to the attention of a staff person at Reddit. This process is considered a

manual review by the corporation, and as such it does not constitute something similar to The Plebeian Algorithm. In order for Reddit to implement a Plebeian Algorithm, it must ensure that the process of the determination of misinformation remains in the hands of the user base.

Although Reddit currently appears to be a democratic system, it is in reality more of a fiefdom.[31] For example, in 2013 the r/FindBostonBombers Subreddit slandered the Brown family by connecting them with the 2013 attack on the Boston Marathon, at the direction of the moderators of the Subreddit.[34] Examples like this resonate throughout Reddit through incidents such as "the Fappening", where nude photographs were released to the public unbeknownst to victims. Incidents like these make apparent the crux of the fundamental issue with Reddit: the moderators. This promotes content moderation by a few elite members of communities, instead of by the members of the said community on the whole.

It should be noted that Reddit is a platform which is built on a sense of anonymity. Users are not required to add their personal email addresses or their real names. As is the case in all of the implementations of the Plebeian Algorithm, it is critical that the social media company critically analyzes the existing market served, and existing qualities which users may be drawn to. An implementation of the Plebeian Algorithm on Reddit should preserve user anonymity, and should still not require the use of personal emails or real full names.

4.4.2 4chan

The 4chan algorithm is similar to that of Reddit; it uses a system whereby the audience determines whether or not content is viewable to its users. In contrast to Reddit, 4chan uses an ephemeral system for its content.[9] 4chan is also divided into several boards which encapsulate distinct topics of conversation. Furthermore, any content can be posted on the /b/ board, as this board's topic is described as "Random".[9] A second critical aspect of the 4chan algorithm is the notion of anonymity. 4chan encourages its user base to remain anonymous through their posts. Over 90% of posts and comments on /b/, the most popular board on 4chan, are anonymous.[9]

4chan is the algorithm which is nearest to the proposed Plebeian Algorithm, however there are subtle yet notable differences. The Plebeian Algorithm does not incorporate any notions of anonymity nor ephemerality. Content must be both traceable and permanently recorded. This will help assure that the goals of the social media companies at-large (which often differ from the goals of 4chan) remain consistent. Keeping the goals consistent for each individual social media platform will be essential to ensure that the users of the platform remain loyal, while gaining the added benefits which the Plebeian Algorithm offers.

In order for 4chan's algorithm to become a Plebeian Algorithm, it should remove its emphemerality. This would be essential to ensure that content has the time to undergo the process. Content posted on 4chan's /b/ often lasts less than 1 minute.[9] As such, the Plebeian Algorithm would not have the time to undergo both of the two phases (the Flag and Jury Phases), a critical step that is necessary to the algorithm's democratic approach.

4.4.3 Facebook

Facebook is a valuable selection, demonstrating a powerful social media platform and a tailored user experience; its popularity making it useful for analysis. While the Facebook company is not wholly transparent,[8] that company has announced highly favours personalized content of users (e.g., posts from close friends and private groups) to that of public groups and pages to which the user likes and follows.[35,36] This presented a limitation for the analysis of Facebook data as ideally, data had to be collected through pre-selected datasets for quantitative analysis.[24]

There are many differences between the Facebook algorithm and the Plebeian Algorithm. The method employed by Facebook, particularly during the COVID-19 pandemic, aims to combat the spread of misinformation is based on neural networks trained to search for key terms in textual elements (including posts, comments, and statuses). It should also be noted that Facebook's algorithm includes a large amount of human work, which is easily biased. As has been stated in prior sections, there are several issues with this method of misinformation censorship; most notably, the Facebook algorithm is limited in scope to a specific subset of topics of misinformation. Algorithms of this nature will detect specific key terms such as "COVID" included in the text of a post to provide additional information and resources for viewers; these algorithms will also provide information to the user sharing the post before publicizing the post. On the contrary, the Plebeian Algorithm is proactive in nature: it is universally applicable to all forms of misinformation, and works to combat infodemics before they become widespread. As stated previously, infodemics of misinformation have led to the problem of pandemics to become exacerbated and thus harder to control for public health workers. By implementing the Plebeian Algorithm, public health will be greatly improved, especially concerning future pandemics, as potentially dangerous misrepresentations or falsehoods about the situation will be contained to a smaller percentage of the populace and thus ensure that reliable and trustworthy information is more accessible and widespread. The Plebeian Algorithm also requires less maintenance by developers, actively running automatically without the requirement of hard-coding key terms to flag.

In order for Facebook to implement a Plebeian Algorithm, a high degree of planning would be required. Since Facebook is the most prevalent social media platform, a gradual implementation based on a rolling basis is recommended. AB testing should be used to ensure a smooth and successful implementation. Facebook should automate and democratize their home page algorithm to implement a Plebeian Algorithm for its service.

4.4.4 YouTube

Due to the inherent difficulty in performing visual sentiment analysis for videos, comments of YouTube videos were analyzed. This does not give a complete picture of the YouTube algorithm, which attempts to keep users engaged longer on the website by presenting a tailored feed; the end-goal being that the algorithm can predict videos the user would like to watch before they search.[37] This algorithm looks at a range of user data including watch time, closing a video tab, the user's interests, freshness, and user interactions with the video.[37] This algorithm has proven highly effective at finding and distributing viral content.

By aligning with the user's sentiment, the algorithm can effectively produce

more positive comments as seen in Figure 8. A sentiment filter used by YouTube includes the removal of videos that do not meet advertiser guidelines.[38] The main difference between YouTube's current content moderation approach and a Plebeian Algorithm's implementation of content moderation is the democratic aspect of content removal. This is made incredibly clear with many of YouTube's controversies within their community revolving around a lack of communication and censorship of larger creators.[39]

The moderation system of YouTube is already a form of Plebeian Algorithm with users being able to like and dislike comments or videos, in addition to reporting them if they are unwanted. The main disconnect between this and the Plebeian Algorithm is that when a comment or video is reported there is no public jury phase where the community decides if it stays. This has become clear with YouTube's controversies within the community revolving around issues such as the lack of communication and censorship of YouTube influencer Logan Paul. Should YouTube implement the Plebeian Algorithm, a Jury Phase is required after content is reported and prior to its removal process. It should also be noted that YouTube's jurors are not a random distribution. The moderation algorithms are programmed by humans, and as such, it is extremely difficult to ensure that the correct decisions are consistently being made. Artificial intelligence forms the basis of the YouTube algorithm, but the Plebeian jury is replaced with a judge, who may be easily persuaded or hold personal biases. The wisdom of the crowd phenomenon (q.v.) plays a huge role in the use of the jury for the Plebeian Algorithm.

4.4.5 Parler

Parler uses a more typical algorithm. It limits posts to one thousand characters, and circulates them to the user base at-large. Thus, unlike Reddit and 4chan, there are no communities in which content is posted on Parler. Parler was founded as a promoter of the freedom of speech, and as such, its user base is highly concerned with a lack of censorship on their posts.[10]

Although this may at face value appear to be in direct opposition to the implementation of any algorithm, it is critical to note that the Plebeian Algorithm ensures that any and all decisions regarding the verity of information remain in the users' hands. Parler would still benefit from implementing a Plebeian Algorithm, as it would preserve Parler's ultimate goal (to promote freedom of expression) while limiting the spread of misinformation.

For Parler to implement a Plebeian Algorithm, it must implement both the Flag and Jury Phases of the Plebeian Algorithm. Notably the preservation of the freedom of expression on the platform must be ensured above all. This will ensure that the user base remains loyal and supportive of the change, and does not boycott Parler or switch to a new social media platform (as they have already migrated from Twitter). The Parler user base is notably precarious, and it must ensure that the user base remains loyal to the platform. This should be done through proper marketing of the transition, which is to be discussed later.

4.4.6 Twitter

Finally, Twitter uses a similar algorithm (in opposition still to Reddit and 4chan) in that posts and content are released to the user base at-large. The techniques of

this algorithm particularly means that misinformation is more likely to spread on Twitter (and Parler) compared to other platforms (e.g. Reddit and 4chan). The large user base on Twitter, as well as the widespread availability of data must be taken into account, as it will be crucial that the culture and atmosphere of Twitter are maintained to ensure that the user base remains content with any algorithmic changes. Twitter's executives most likely would be interested in increasing their reach by attempting to regain the trust of those who migrated to Parler. These individuals are highly concerned with a decrease in censorship and an increase in the freedom of expression. They believe that the social media platform should remain separate from the process of promotion and demotion of content.[10]

In order for Twitter to implement a Plebeian Algorithm, it must attempt to promote the freedom of expression and a decrease in censorship, while also maintaining their reliability. This is done using the Plebeian Algorithm, which takes advantage of both of the concerns. Layperson algorithms have proven effective at curbing the spread of misinformation, and at increasing reliability.[17] According to the definition by Epstein et al., the Plebeian Algorithm would be classified as a type of layperson algorithm. Twitter would need to place a level of trust in the layperson to provide the user base with liberty, while maintaining truth in content posted.

4.5 Condorcet's Jury Theorem

As was researched in the eighteenth century by the Marquis de Condorcet, the Condorcet's Jury Theorem[40] clearly justifies the need for the Jury Phase in the Plebeian Algorithm. The theorem describes the behaviour of a larger number of individuals selected to sit on a jury, to judge the crimes of another individual. Proposed by Condorcet (and later proven by numerous mathematicians and statisticians in the late twentieth century), the theorem explains that two scenarios may unfold when attempting to determine the truth by means of polling a sample of the population.[40] Firstly, if the sample's understanding of the topic is poor, their judgement will not be certain. In this situation, the optimal sample size would be a single individual, as increasing the number of jurors will only increase the uncertainty.[40]

However, a jury implemented in the Plebeian Algorithm should take Condorcet's Jury Theorem into account, by ensuring that the jury falls into the second scenario. The second type of jury would occur when the jury's knowledge of the subject is relatively high, or is perceived as relatively high.[40] As such, an optimal Plebeian model would be passive, instead of aggressive, in it's UI/UX. It should be ensured that acting as a juror is entirely optional, and is opt-in instead of opt-out. The user interface should be minimal, in order to ensure that the public reception of the implementation of the Plebeian Algorithm is positive. Although this will likely decrease the percentage of the sample who opt-in to act as jurors, consistence will be achieved due to the positive reception of the implementation of the algorithm. An ideal Plebeian Algorithm implementation, in order to secure the second subset of juries defied by Condorcet's Jury Theorem, may go unnoticed for the average user.

Assuming that an implementation of the Plebeian Algorithm is able to secure its jury into the latter jury type, it would secure the wisdom of the crowd. Increasing the sample selected as potential jurors will increase the certainty. This phenomenon has been described as "wisdom of the crowd".[41] As the sample size increases, the certainty of the decision which the jury comes to also increases. Thus, taking a

sample size of 1% has a higher possibility of accidentally selecting a group of the most extreme individuals, compared to randomly selecting a sample size of 10%.

4.6 Eradication vs. Containment

One of the benefits of the Plebeian Algorithm in comparison to the status quo is the difference between eradication and containment. The algorithms of the current system tend to use an eradication approach. They view the issue of the spread of misinformation with a narrow perspective,[42] and as such, they tend to implement a "Point-And-Shoot" Algorithm. With this system, media companies determine which posts contain misinformation and eradicate them on a case-by-case basis. For example, many social media companies utilize a COVID-19 key term search, and flagging posts that contain them. They then link to a government website with information on the pandemic to the post.

This is in stark contrast to the Plebeian Algorithm, which takes a containment-based approach to the spread of misinformation. It should be noted that the technology to remove all instances of misinformation does not exist.[42] Instead, it is critical that the algorithm detects as many cases of misinformation as possible and to bring the rest to the broad public. This essentially "pops" any filter bubbles and echo chambers.[26] It allows for positive discussion from the community, which tends to lead towards a decrease in misinformation.[5]

4.7 Reduced Censorship

Another massive benefit of the Plebeian Algorithm is the reduction of censorship. With regards to the COVID-19 pandemic, a majority of misinformed posts has been spread by those with politically-right ideologies, or Republicans.[43] 48% of those with U.S. Republicans believe that SARS-CoV-2 is no more dangerous than the common influenza[43] (compared to 25% of U.S. Democrats[43]), and 42% of those with Republicans believe that hydroxychloroquine – a treatment for malaria – is an effective treatment for SARS-CoV-2[43] (compared to 5% of Democrats).[43] Furthermore, those with political-right ideologies tend to be more concerned with the preservation of the freedoms of speech and expression. Thus, it is evident that we must preserve these freedoms for any algorithmic change to be effective. The Plebeian Algorithm goes further than this: it works to increase the rights of the individuals with respect to cross region community matching. freedom of expression. Individuals have the right to post and speak as they please and promote the spread of the information they deem to be pertinent. Additionally, they have the right to decide what content they want to see on the platform and what they don't. These benefits will help to ensure that the public reacts in a positive light to the change. Implementing a Plebeian Algorithm is a net positive; it is a positive change for both the containment of infodemics and the promotion of freedom of expression among social media users. Furthermore, under the Plebeian Algorithm, social media companies are still permitted to analyze user activity according to their privacy policies to provide appropriate advertisements tailored to the user. This will ensure that the revenues for social media companies will not be reduced in the process.

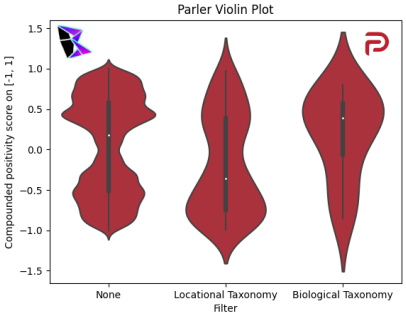

Figure 13: Viral Naming Conventions on Parler: Violin Plot

4.8 Viral Naming Conventions

One sub-topic explored herein is viral naming conventions and the connection between the name used to describe COVID-19 in relation to the the level of verity in social media posts. For the purpose of this analysis, only posts on COVID-19 were considered; thus, social media platforms for which the data used herein was collected before 2020 were not analyzed (i.e. Facebook and YouTube). Parler was examined at length since it uses a relatively standard social media algorithm, comparable to that of Twitter. A pickle file of Parler data, filtered to COVID-19, was generated from the Parler dataset.[10] Additional filters were applied to the pickle file, as are described below.

To simplify the analysis, three categories of Parleys were created on which analysis was performed separately. First, all posts mentioning COVID-19 using any naming conventions were gathered. The posts were collected using no additional filters, labelled "None". Second, from the COVID-19 Parleys, a filter was applied to gather all Parleys containing viral names referring to locations including, but are not limited to, "Wuhan Virus", "China Virus", and "Indian Variant". All of these terms have been described by the United States Center for Disease Control (CDC) to potentially propagate misinformation and xenophobia.[44,45] This filter was termed "Locational Taxonomy".[46] The final filter, "Biological Taxonomy",[46] refers to the biological names for COVID-19, or officially approved names by the World Health Organization, including, but are not limited to, "SARS-CoV-2", "Alpha Variant" and "B.1.617". This nomenclature is used and promoted by the CDC[45] to limit xenophobia. As such, it was hypothesized that Parleys using these terms would be less likely to be misinformative and more likely to have positive sentiment.[47]

Sentiment analysis was performed on all three of the sets of filtered data, and the results were plotted as a violin plot in Figure 13. Each subplot portrays the three filters as discrete categories along the x-axis (i.e. "None", "Locational Taxonomy", and "Biological Taxonomy"), against the compounded polarity score on $[-1, 1]$. Each subplot visualizes a KDE that is rotated vertically for ease of visualization and a pictorial representation of the median, mean, first quartile and third quartile.

This visualization provides exceptionally relevant results. The data filtered to COVID-19 at-large was similar to the KDEs plotted for all of the social media algorithms discussed in the Results section, demonstrating a precise bimodal distribution with a positive and negative peak, and a neutral trough. The violin plots for the locational and biological taxonomies verified the hypothesis. The locational taxonomy filter showed strong negative sentiment, implying a higher likelihood of

misinformation. In contrast, the biological taxonomy filter showed a strong positive sentiment, implying a greater degree of verity.

It is critical to note that the findings are not limited to COVID-19. Similar findings were discovered (not pertaining to social media) relating to the 2009 H1N1/09 pandemic,[48,49] the 1918 Spanish Flu pandemic,[16] among others,[50,51] for which there were concerns surrounding xenophobic viral nomenclature. It is also pertinent to discuss the specific limitation surrounding the use of COVID-19 data for this analysis. It is often difficult for populations to alter their vocabulary to change the reference of a xenophobic initial name, to an accurate descriptor.[52] Specifically, with regards to COVID-19 variants of concern (VoCs), many scientific sources still note the location of discovery of the VoC. This brings two significant points to the forefront of discussion: first, national health agencies need to provide precise and non-xenophobic nomenclature from the onset of pandemics, and second, the use of locational taxonomy should not automatically flag a post (i.e. it should be flagged through sentimental analysis solely). The Plebeian Algorithm assists in this analysis, as it does not consider specific search terms, but rather pure sentiment. This handles issues surrounding truthful posts containing locational taxonomy. The lack of consensus among the scientific community should be noted with regard to the potential benefits and drawbacks of using locational taxonomy.[16,47]

4.9 Geolocation

There exists a critical connection between the virtual and physical worlds as it pertains to the spread of misinformation and various consequences therefrom. Several studies have been conducted hereupon. One fundamental limitation posed by the use of social media platforms to track the spread of misinformation is the inability to deal with the spread of misinformation in more personal settings (e.g. face-to-face interactions, video conferencing, and direct messaging). Thus, a thorough study of the translation of misinformation from social media platforms to real-world phenomena will be conducted.

Myriad studies have been conducted surrounding infodemics.[53–55] This includes the correlation between geolocation of social media connections and various social determinants (e.g. race, sex, and socioeconomic status),[53] and a study[56] determining optimal methods of geolocation on social media. Two additional studies discussed the sociological consequences of geolocation in the context of social media, namely the detection and reduction of youth cannabis consumption[54] and the applications of geolocation to urban planning.[55]

Furthermore, studies suggest that there exists a strong correlation between trends on social media and events such as COVID-19 infections.[13,53] Evidently, any change on social media will have real-world impacts. Thus, it is apparent that a reduction in the amount of misinforming content in a social media user's home page corresponds with a reduction in the likelihood that they will propagate misinformative statements when having in-person conversations. Successful implementation of the Plebeian Algorithm will limit the spread of misinformation on social media platforms and in the lives of their users.

4.10 Public Reaction

Skeptics of the Plebeian Algorithm might be concerned that such a massive alteration of the social media algorithm will incite hesitancy from the public. Whether this hesitancy takes the form of negative feedback or boycotting, it is legitimate and must be dealt with. Many will point to the 4chan platform as a negative example of an algorithm that offers user discretion regarding the promotion of content instead of a corporate algorithm.

4.10.1 Marketing

This paper will firstly argue that the major difference between the two strategies lies within the realm of marketing. Marketing is a critical aspect of any social media company, especially when undergoing massive changes. In fact, some broad-scale social changes require marketing strategies.[57] Companies must ensure that the Plebeian Algorithm is adapted to meet the specific needs and goals of the social media company and its respective user base. For this reason, the Plebeian Algorithm is simply a suggested implementation, with a footnote that the algorithm must be highly adapted to the unique situation. Every social media company has varying objectives, such as Facebook's aim to connect friends, Reddit's goal to create conversations between like-minded individuals, and Parler's goal of preserving freedom of expression.

An effective marketing strategy for the transition to the Plebeian Algorithm ensures that users are aware that the overall atmosphere of the social media platform will not be altered. Promotion of the current atmosphere must take priority, lest the change face backlash by users. There is a potential, should improper marketing be implemented, that overly-moderated individuals may leave the social media platform, leaving those with more extreme (and often misinformed) views to take over the widespread content of the platform. However, adequate marketing that emphasizes the static nature of the culture and social atmosphere of the platform during the transition alleviates this concern.

4.10.2 Feedback of Current Algorithms

Secondly, this paper will discuss the feedback on the current algorithms as provided by the community. This feedback consists of discussions on social media platforms about each platform's respective algorithm. An analysis of pre-selected opinion pieces was performed.[35,36,58–64] These opinion pieces were sourced from well-known news or magazine sources, discussing the various social media platforms analyzed herein.

On the whole, there is an immense desire for social media platforms to be more democratic in their algorithm. It is also widely believed among many social media users that, in order to improve algorithms, companies should implement a more transparent algorithm. Currently, algorithms vary widely and the functionality of most are not publicly available information. Changes improving transparency tend towards positive user feedback on the platform.

It is also critical to note that for any implementation of the Plebeian Algorithm, a post must exceed a popularity threshold to be flagged in the Flag Phase. It is essential for the social media platforms to adapt their current algorithm to the

determination of this popularity threshold. The goal of most current algorithms is to show users popular content which they may enjoy based on past interests. This can be done through a plethora of metrics, including likes, views, comments, recency of the post (termed "freshness"),[37] and more. For example, the Twitter algorithm tends to prioritize the number of comments, whereas the YouTube algorithm prioritizes freshness.

4.11 Implementation

The second concern of a potential implementing platform of the Plebeian Algorithm would be the technological requirements of the implementation, including storage and processing power required to conduct the Plebeian Algorithm on their millions of posts. Furthermore, this application of the Plebeian Algorithm would need to be a continuous process, ensuring that the algorithm continually updates when new comments are added to a post. As has been shown herein, the inclusion of comments increases the level of detail. All of the data analyses visualized herein included comments, and the text of the original post. As such, the computational power required appears to be great. There are, however, many alterations that can be made to the Plebeian Algorithm to reduce computation costs.

Firstly, the Plebeian Algorithm does not need to be updated with the post of every new comment. It can be performed on intervals, whereby a subsection of posts is checked for new comments at every time interval. These new comments (and only the new comments) are then sent through sentiment analysis. In terms of data storage, it may prove useful for the social media platform to store a single additional byte of data for each post. The bit of highest significance, referred to as the "Flag of Need Determination", represented as ϕ_{det}, can be defined using the following equation:

$$\phi_{\text{det}}(x) = \left\lfloor \frac{1}{2}\text{sgn}(n_{\text{det}}(x)) + 1 \right\rfloor \tag{2}$$

such that,

$$n_{\text{det}}(x) = \sum_{i=0}^{N} \begin{cases} v_{\text{act}} p(x_i) - v_{\text{thres}} & \text{if } \neg\nu(x) \\ 0 & \text{else} \end{cases} \tag{3}$$

where n_{det} represents the value of determination (which is not scaled), x represents a thread, x_i represents a specific comment or post within a thread, ν is a Boolean function returning a high value if the comment is new and low if it has been analyzed, sgn represents the signum function, v_{act} and v_{thres} represent the actual and threshold popularity of a thread in number of views and N as the number of posts or comments in the thread.

If high, the post or thread can be safely skipped by the algorithm. If low, the post or thread will be analyzed in order to ensure that no misinformation goes undetected. The remaining seven bits of the data represent the sentiment of the entire thread, represented using β_N, where N is the number of comments in the post, excluding the original post. These bits can be calculated using the following equations:

$$\beta_N = \frac{32}{N} \sum_{i=0}^{N} p(x_i) \forall N \in \mathbb{Z}^+ \tag{4}$$

In some circumstances, it may be more computationally convenient to calculate β_N recursively, which may be done using the following:

$$\beta_N = \frac{1}{2}\beta_{N-1} + 16p(x_i) \tag{5}$$

These equations demonstrate that a byte can be associated with each thread to decrease the processing requirements to execute the Plebeian Algorithm on a large scale.

It should also be noted that the Plebeian Algorithm is a machine learning model. It can be built to work in tandem with existing machine learning algorithms, thus decreasing the computing power required. Data storage is minimized using the one-byte storage method described above. As is the case with all neural networks, the Plebeian Algorithm's Flag Phase will increase in accuracy over time by manipulating the string data as a validation set. Thus, the neural network will improve in accuracy over time. Due to time and resource limitations, the paper used the Valence Aware Dictionary and sEntiment Reasoner (VADER), however in order to increase Flag Phase precision over time, it is recommended that platforms implement the VADER sentiment analysis tool initially, but built on it in order to adapt to the specific lexicon of the social media platform, at the period of time. This accounts for minor differences in various social media algorithms and for lexical changes over time.

It is critical that a public release of the Plebeian Algorithm should be done through a process of AB testing. In order to efficiently fix any inevitable bugs present in the implementation of the algorithm (including any potential philosophical issues surrounding a specific realization/implementation), AB testing will be vital in the assurance that users consuming media under the new algorithm remain loyal to the brand, and minimize any potential negative impacts. It will allow user feedback to be gathered for the small subsection of users presented the Plebeian Algorithm implementation.

4.12 Limitations

While the Plebeian Algorithm is a great replacement for the current attempts by social media platforms to reduce the spread of misinformation, it is limited by several key factors. First, as stated earlier, the algorithm was only is confirmed applicable for strictly text-based social media platforms and posts. Thus, the moderation of videos or images are outside of the scope of its usage. Second, private sources of media such as chat rooms and servers are not within the scope of the algorithm and thus the algorithm is limited to public communication media. Thirdly, the determination of a popularity threshold can be problematic. On Twitter, for example, a significant number of Re-Tweets are done passively (i.e. they are not done for the express purpose of sharing with others, but are done subconsciously by the user). Passive sharing may cause issues in the determination of whether a piece of content meets a popularity threshold. Finally, it is limited in the sense that it cannot determine what is misinformation at an instantaneous time selection, and as such misinformation cannot be extracted from the algorithm at any time.

Conclusions

The implications of this research are significant as to provide social media platforms with a new flagging method that utilizes sentiment analysis. This will be critical in the detection and prevention of infodemics and utilizing a democratic approach that gives the power to the social media user to ultimately decide what content should be on the platform, based on accuracy. The Plebeian Algorithm directly reduces political polarization and extremist ideas, which create a divide among users and improve cooperation on resolving key issues and problems plaguing humanity and restore the trust between the public and experts.

Additionally, it is predicted that this will result in more reliable social media platforms, leading to an overall reduction of ignorance and misinformed opinions among users. Finally, the model created will lead to users expressing themselves without concern of the political viewpoint of the social media platform. Inherently, this also minimizes the impact of external biases, such as political climate, as those who vote will be completely random and anonymous.

Many future areas of research concerning the domain remain un-analyzed. These topics include, but are not limited to:

- Conducting a study on the utilization of the Plebeian Algorithm on a selection of social media platforms and detecting the amount of misinformation over time after its implementation (i.e. a real-world tested example) which would then be compared to current methods used such as the aforementioned "Point-And-Shoot" Algorithm;
- Creating a type of sentiment analysis for graphical content which could examine the emotion within an image to determine if it could be misinformation (e.g. Snapchat, Instagram and TikTok);[65,66]
- Determining the spread of misinformation correlated with the spread of viruses – this could be useful in pre-determining locations (and users by extension) who are at higher risk of being exposed to or expounding misinformation;
- Exploring the applicability of the Plebeian Algorithm in surveillance contexts, including for criminal investigations, employee onboarding and healthcare;[67–70]
- Analysing the spread of misinformation through online vendors such as Amazon or eBay. In particular recent audits of Amazon (as of 2021) show a dangerous disregard for reliable information; for example, presenting vaccine misinformation books along with well cited vaccine information books in generic searches for vaccine information.;[71–75]
- Applying models of higher sophistication for data analysis and visualization (which requires access to more in-depth data), including tf-idf measures[76] and Levenshtein distances[77] among others;[78]
- Examining the optimal method of implementation and integration for the Plebeian Algorithm with various existing networking systems and infrastructures;
- Continuing analysis of data collected to corroborate to prior studies on behavioural impacts of the sentiment of informative posts on social media;

- Analyzing the role of corporate social media platforms (i.e. Slack) in the dissemination of misinformation, especially in private chat channels;
- Examining the misinformation containment models using juries, including the jury system implemented by Wikipedia;
- Analyzing the rise of audio-form content, including podcasts, Clubhouse and Spotify Greenroom audio-chat rooms, for the potential spread of misinformation – many of these media are becoming increasingly influential sources of news and information for many;[79] and,
- Exploring the connection between location-based social media apps (such as Foursquare), at the spread of geographic misinformation.[80]

COVID-19 has had significant impacts upon the modern society. Optimists hoped these impacts would prove to unite a polarized world in the spirit of cooperation and global security. While this has happened, their hopeful unity to the political schism has not. The Plebeian Algorithm is not a vaccine for an infodemic; however, it is a treatment to help curb and prevent the virus of misinformation from continuing to spread and grow out of control. This has the critical side-effect of putting the power back in the hands of the people and removing the potentially-domination of a single entity (e.g. a social media company) who may be swayed by external forces when deciding if content should be removed. All in all, it is recommended that social media executives consider the implementation of a variation upon the Plebeian Algorithm, explicitly modified to adapt to the specifics of the platform. This will help curb misinformation both with regards to the COVID-19 infodemic as well as to prevent future infodemics.

Acknowledgements

The authors would like to acknowledge the assistance in ideation from Anish R. Verma from STEM Fellowship. In addition, the authors are grateful for the assistance provided by Vinayak Nair concerning refinement of arguments from a data science perspective. Sponsors for the Undergraduate Big Data Challenge 2021, including JMIR Publications, Roche, SAS, Canadian Science Publishing, Digital Science, and Overleaf, made it possible for this research to be conducted.

References

[1] Soucheray S. Poll: 1 of 4 Americans will refuse COVID-19 vaccine https://www.cidrap.umn.edu/news-perspective/2021/03/poll-1-4-americans\-will-refuse-covid-19-vaccine 2021. Published by Center for Infectious Disease Research and Policy at the Univeristy of Minnesota.

[2] A Muacevic, J Alder. Social Media Use and Its Connection to Mental Health: A Systematic Review *Cureus.* 2020;12.

[3] Kim HHS. The impact of online social networking on adolescent psychological well-being (WB): a population-level analysis of Korean school-aged children https://www.tandfonline.com/doi/full/10.1080/02673843.2016.1197135 2016.

[4] Watson A. Share of adults who use social media as a source of news in selected countries worldwide as of February 2020 https://www.statista.com/statistics/718019/social-media-news-source/ 2020.

[5] Walsh D. Neutral Isn't Neutral: An Analysis of Misinformation and Sentiment in the Wake of the Capitol Riots Master's thesisWest Virginia University 2021.

[6] Wardle C, Derakhshan H. Information Disorder: Toward an interdisciplinary framework for research and policy making https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c 2017.

[7] Bradd S. Infodemic https://www.who.int/health-topics/infodemic#tab=tab_1 2020.

[8] al. K. Hazelwood. Applied Machine Learning at Facebook:A Datacenter Infrastructure Perspective https://research.fb.com/wp-content/uploads/2017/12/hpca-2018-facebook.pdf 2017.

[9] Bernstein M et al. 4chan and /b/: An Analysis of Anonymity and Ephemerality in a Large Online Community https://www.researchgate.net/publication/221297869_4chan_and_b_An_Analysis_of_Anonymity_and_Ephemerality_in_a_Large_Online_Community 2011.

[10] Aliapoulios M. An Early Look at the Parler Online Social Network https://arxiv.org/pdf/2101.03820.pdf 2021.

[11] Massanari A. Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures *New Media and Society.* 2015;19.

[12] K Kumar, G Geethakumari. Detecting misinformation in online social networks using cognitive psychology *Human-centric Computing and Information Science.* 2014;4.

[13] Stephens M. A geospatial infodemic: Mapping Twitter conspiracy theories of COVID-19 *Dialogues in Human Geography.* 2020;10.

[14] Fernandez M, Belloghin A, Cantador I. Analysing the Effect of Recommendation Algorithms on the Amplification of Misinformation Print 2021.

[15] Plaza M et al. The use of distributed consensus algorithms to curtail the spread of medical misinformation *International Journal of Academic Medicine.* 2019;5.

[16] Hoppe T. Spanish Flu: When Infectious Disease Names Blur Origins and Stigmatize Those Infected *American Public Health Association.* 2018;108.

[17] Epstein Z, Pennycook G, Rand D. Will the Crowd Game the Algorithm? Using Layperson Judgments to Combat Misinformation on Social Media by Downranking Distrusted Sources https://dl.acm.org/doi/10.1145/3313831.3376232 2020.

[18] Kash K. Covid Vaccine Tweets https://www.kaggle.com/kaushiksuresh147/covidvaccine-tweets 2021.

[19] Carley K, Memon S. CMU-MisCov19: A Novel Twitter Dataset for Characterizing COVID-19 Misinformation https://doi.org/10.5281/zenodo.4024154 2020.

[20] Gruzd A, Mai P. Inoculating against an Infodemic: A Canada-wide COVID-19 News, Social Media, and Misinformation Survey https://doi.org/10.5683/SP2/JLULYA 2020.

[21] Ambalina L. . https://lionbridge.ai/datasets/top-20-twitter-datasets-for-natural\-language-processing-and-machine\-learning/ 2019.

[22] Palachy S. Twitter Datasets https://github.com/shaypal5/awesome-twitter-data 2020.

[23] Tankovska H. Number of Monthly Active Facebook Users Worldwide as of 1st Quarter 2021 https://www.statista.com/statistics/264810/\number-of-monthly-active-facebook\-users-worldwide/ 2021.

[24] Span J. FacebookR Comments https://github.com/jerryspan/FacebookR 2017.

[25] Mitchell J. Trending YouTube Video Statistics and Comments https://www.kaggle.com/datasnaek/youtube 2017.

[26] Rhodes S. Filter Bubbles, Echo Chambers and Fake News: How Social Media Conditions Individuals to be Less Critical of Political Misinformation https://www.openicpsr.org/openicpsr/project/135024/version/V2/view 2021.

[27] Wollebaek D, Karlsen R, Steen-Johnsen K, Enjolras B. Anger Fear and Echo Chambers: The Emotional Basis for Online Behavior *Sage.* 2019;5.

[28] E Pariser. The filter bubble: What the Internet is hiding from you. https://hci.stanford.edu/courses/cs047n/readings/The_Filter_Bubble.pdf 2011.

[29] Frerichs RR. Simple Random Sampling https://www.ph.ucla.edu/epi/rapidsurveys/RScourse/RSbook_ch3.pdf 2008.

[30] Rosenburg H, Syed S, Rezaie S. The Twitter pandemic: The critical role of Twitter in the dissemination of medical information and misinformation during the COVID-19 pandemic *Canadian Journal of Emergency Medicine.* 2020;22:418–421.

[31] Auerbach D. Reddit Scandals: Does Reddit Have a Transparency Problem? https://slate.com/technology/2014/10/reddit-scandals-does-the-site-\have-a-transparency-problem.html 2014.

[32] DeVito M, Gergle D, Bernholtz J. Algorithms ruin everything: RIPTwitter, Folk Theories, and Resistance to Algorithmic Change in Social Media *HCI and Collective Action.* 2017.

[33] University Cornell. Comparative Study of HITS and PageRank Link based Ranking Algorithms https://blogs.cornell.edu/info2040/2015/10/27/comparative-study-of-hits-and-\pagerank-link-based-ranking-\algorithms/ 2015.

[34] Mcintyre M. Relational Agency, Networked Technology, and the Social Media Aftermath of the Boston Marathon Bombing Master's thesisUniversity of South Florida 2015.

[35] Hutchinson A. The Pros and Cons of Facebook's Coming News Feed Changes - from a Page Perspective https://www.socialmediatoday.com/news/the-pros-and-cons-of-facebooks-\coming-news-feed-changes-from-a-\page-pers/514776/ 2018.

[36] Cooper P. How the Facebook Algorithm Works in 2021 and How to Make it Work for You https://blog.hootsuite.com/facebook-algorithm/ 2021.

[37] Covington P, Adams J, Sargin E. Deep Neural Networks for YouTube Recommendations https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45530.pdf 2016.

[38] Alphabet Inc. YouTube Terms of Service https://www.youtube.com/static?template=terms 2021.

[39] Southerton C et al. Restricted modes: Social media, content classification and LGBTQ sexual citizenship *New Media & Society.* 2020;23:920–938.

[40] Austen-Smith D, B Jeffrey. Information Aggregation, Rationality and the Condorcet Jury Theorem *The American Polticial Science Review.* 1996;90:34–45.

[41] Yi et al.. The Wisdom of the Crowd in Combinatorial Problems *Cognitive Science.* 2012;36:452–470.

[42] Ghoshal AK, S Das, N Das. Influence of community structure on misinformation containment in online social networks *Elsevier B. V..* 2020.

[43] Cox D. Conspiracy theories, misinformation, COVID-19, and the 2020 election tech. rep.Survey Center on American Life 2020.

[44] Abdool Karim S, Oliveira T, Loots G. Appropriate names for COVID-19 variants *Science.* 2021;371.

[45] Center For Disease Control. Emerging Infectious Diseases: Scientific Nomenclature https://wwwnc.cdc.gov/eid/page/scientific-nomenclature 2014.

[46] Hull R, Rima B. Virus taxonomy and classification: naming of virus species *Arch. Virol..* 2020;165:2733–2736.

[47] Masters-Waage T, Jha N, Reb J. COVID-19, Coronavirus, Wuhan Virus, or China Virus? Understanding How to Do No Harm When Naming an Infectious Disease *Front. Psychol..* 2020;11:1–10.

[48] Vigsø O. Naming is Framing: Swine Flu, New Flu, and A(H1N1) *Observatorio.* 2010;4.

[49] McCauley M, Minsky S, Viswanath K. The H1N1 pandemic: media frames, stigmatization and coping *BMC Public Health.* 2013;13.

[50] Sell T, Hosangadi D, Trotochaud M. Misinformation and the US Ebola communication crisis: analyzing the veracity and content of social media messages related to a fear-inducing infectious disease outbreak *BMC Public Health.* 2020;20.

[51] Singh RP et al. The naming of Potato virus Y strains infecting potato *Arch. Virol..* 2007;153:1–13.

[52] Mallapaty S. Should Virus-Naming Rules Change During a Pandemic? *Springer Nature.* 2020;584:19–20.

[53] Baucom E et al. Mirroring the Real World in Social Media: Twitter Geolocation, and Sentiment Analysis *ACM.* 2013.

[54] Chung T. Innovative Routes for Enhancing Adolescent Marijuana Treatment: Interplay of Peer Influence Across Social Media and Geolocation *Curr. Addict. Rep..* 2016:221–229.

[55] Milusheva S. Applying machine learning and geolocation techniques to social media data (Twitter) to develop a resource for urban planning *PLoS ONE.* 2021;16.

[56] Williams E, Grey J, Dixon B. Improving geolocation of social media posts *Pervasive and Mobile Computing.* 2016;36:68–79.

[57] Kolter P, Zaltman G. Social Marketing: An Approach to Planned Social Change *Journal of Marketing.* 1971;35:3–12.

[58] Mayfield D. Social Media Algorithms 2021: Updates and Tips by Platform https://storychief.io/blog/social-media-algorithms-updates-tips 2021.

[59] Settlage B. Pros and Cons of the Social Media Algorithm Age https://www.amplimark.com/pros-and-cons-of-the-social-media-\algorithm-age/ 2021.

[60] Company On The Maps! Digital Marketing. Disadvantages Of The Facebook Algorithm And How You Can Workaround It. https://onthemaps.com/disadvantages-of-the-facebook-\algorithm-and-how-you-can-\workaround-it/ 2020.

[61] Thottam I. Pros and Cons of Twitter's New Algorithmic Timeline https://www.pastemagazine.com/tech/twitter/pros-and-cons-of-twitters-new-\algorithmic-timeline/ 2016.

[62] Curvelo R. The Pros and Cons of using Twitter as part of your Digital Marketing Strategy https://www.matrixinternet.ie/the-pros-and-cons-of-twitter/ 2020.

[63] Casper H. Everything You Need To Know About Twitter's Timeline Algorithm https://clicktotweet.com/blog/everything-you-need-to-know-about-\twitter-timeline-algorithm/ 2017.

[64] Beasing D. Does Facebook's New Algorithm Add Up for Radio? *Radio World.* 2018;42:22.

[65] Agung N, Darma G. Opportunities and Challenges of Instagram Algorithm in Improving Competitive Advantage *International Journal of Innovative Science and Research Technology.* 2019;4.

[66] Faddoul M. Why is TikTok creating filter bubbles based on your race? https://www.ischool.berkeley.edu/news/2020/\alumnus-marc-faddoul-discovers\-racial-biases-tiktoks-algorithm 2020.

[67] Mateescu A et al. Social Media Surveillance and Law Enforcement http://www.datacivilrights.org/pubs/2015-1027/Social_Media_Surveillance_and_Law_Enforcement.pdf 2015.

[68] Gupta A, Katarya R. Social media based surveillance systems for healthcare using machine learning: A systematic review *J. Biomed. Inform..* 2020.

[69] Bizzi L. Should HR managers allow employees to use social media at work? Behavioral and motivational outcomes of employee blogging https://doi.org/10.1080/09585192.2017.1402359 2017.

[70] Alexander E, Mader D, Mader F. Using Social Media During the Hiring Process: A Comparison Between Recruiters and Job Seekers https://digitalcommons.kennesaw.edu/cgi/viewcontent.cgi?article=1203&context=ama_proceedings 2015.

[71] Shin J, Valente T. Algorithms and Health Misinformation: A Case Study of Vaccine Books on Amazon *Journal of Health Communication.* 2020;25:394–401.

[72] Juneja P, Mitra T. Auditing E-Commerce Platforms for Algorithmically Curated Vaccine Misinformation *CHI Conference on Human Factors in Computing Systems.* 2021.

[73] Smith B, Linden G. Two Decades of Recommender Systems at Amazon.com *IEEE The Test of Time.* 2017.

[74] Krishnamurthy S. A Comparative Analysis of eBay and Amazon https://faculty.washington.edu/sandeep/d/amazonebay.pdf 2004.

[75] Yuan T, Chen Z, Mathieson M. Predicting eBay L isting Conversion http://www.bayimage.com/code/Sigir2011_docs_p1335.pdf 2011.

[76] Aizawa A.. An information-theoretic perspective of tf–idf measures *Information Processing and Management.* 2003.

[77] Andoni A, Onak K. Approximating Edit Distances in Near-Linear Time in *Proceedings of the 41st Annual ACM Symposium on Theory of Computing*(Cambridge, MA):199–204Massachusetts Institute of Technology 2009.

[78] Hong I, Rutherford A, Cebrian M. Social mobilization and polarization can create volatility in COVID-19 pandemic control *Appl. Netw. Sci..* 2021;6.

[79] Alang N. Clubhouse chat beating social media trolls: For better or worse, site offers conversation among the like-minded 2021. Copyright - Copyright 2021 Toronto Star Newspapers Limited. All Rights Reserved; Last updated - 2021-03-06.

[80] Zhao YL, Nie L, Wang X, Chua TS. Personalized Recommendations of Locally Interesting Venues to Tourists via Cross Region Community Matching *ACM Transactions on Intelligent Systems and Technology.* 2013;5:1–26.