Project Proposal
Information Visualization
(CSCI 628)

Ashiqur Rahman
Z-ID: **Z1874484**

16 September 2021

Abstract

Graphical representation of complex data can exhibit interesting perspectives and obscure information. During the Covid-19 pandemic, misinformation on social media was one of the main battlefields for scientists. In this project, we want to explore the change of misinformation topics over time and the possible impact on public perception regarding the vaccine. We expect to find a significant way to communicate this impact of misinformation visually.

Coordinator

Dr. David Koop
Assistant Professor
Department of Computer Science
Northern Illinois University

Contents

1 Introduction

Social media plays a vital role in modern life by communicating, disseminating information, and steering social conversations [25]. These features have more of an impact during a state of crisis [2]. We have seen this during major events like mass shootings, natural disasters, national elections, and even anti-vaccination campaigns [2], [4], [7], [8].

The current pandemic, due to the Coronavirus, created significant dependence on social media. While social media has an essential role in spreading updated information and creating public awareness, misinformation has spread with little to no oversight as well. As soon as the coronavirus emerged, racism, rumor, and fear-mongering started spreading like wildfire on different platforms [14]. This issue had some dire consequences like an overdose of Chloroquine in Nigeria, the suicide of a father of three in India, and the shortage of Hydroxychloroquine for Lupus patients as a direct influence of misinformation circulating the web [20], [22]. Well-intended dissemination of unsubstantiated information can do more harm than good [19]. Also, malicious characters try to take political, financial, or otherwise advantages using misinformation [8], [9]. The World Health Organization (WHO) has recently partnered with major tech giants such as Facebook, Google, LinkedIn, Microsoft, Reddit, and Twitter to fight against misinformation [22]. However, misinformation is still widely available on these platforms.

Although several precautionary techniques like masking, social distancing, and proper sanitization [27] are available, history teaches us that vaccination is the most meaningful strategy to overcome this global pandemic [10]. While some promising vaccines are developed and are accessible around the world [30], it will require a significant percentage of the population to be vaccinated to achieve herd immunity [24]. In comparison, measles requires 95% vaccination, and polio requires 80% vaccination coverage to achieve this collective herd immunity [24]. Unfortunately, Anti-vaxxers and conspiracy theorists are actively creating and disseminating misinformation about vaccination. This misinformation campaign can significantly thwart the effort to reach that goal of herd immunity and put this pandemic behind us [26]. As a result, it is crucial to study misleading information about Covid-19 vaccination and find the best way to combat the anti-vaccine campaigns.

In this project, we plan to analyze Twitter data during this pandemic and find a meaningful way to convey the impact of misinformation and how that changes the opinion of people towards vaccines over time.

2 Background

There is a thin line between fake news and misinformation. When verifiable false news is shared, that is called fake news [6]. On the other hand, misinformation is when false news is shared unintentionally [1], [23]. The mass population can genuinely believe fake news without verifying it and take part in spreading misinformation.

After analyzing 43.3 million tweets, Ferrara E. [16] found that automated social bots are used to disseminate misinformation and political conspiracy theories related to COVID-19. Al-Rakhami and Al-Amri [21] proposed a framework to use six different machine learning algorithms to detect misinformation. They collected the data using Twitter API at the beginning of the pandemic and manually labeled the data to train the models.

Even the vaccination to prevent COVID-19 is being debated, and the misinformation is spread by the opponents of vaccination more frequently compared to the proponents [17]. Although officials are taking steps to handle the misinformation regarding the vaccine [13], the efforts are still falling short [29] to tackle the diverse reasons [18] for the spread of misinformation.

This seemingly unstoppable spread of misinformation can change the objectivity of the population towards vaccination. Sentiment analysis and stance detection can give us a clear overview of the impact on public opinion. Sentiment analysis is one of the major research areas in natural language processing (NLP) and can help us determine the overall perception of the population about any topic. Stance detection is somewhat different than traditional sentiment analysis. While sentiment analysis can detect whether a block of text is positive, negative, or neutral, stance detection can classify someone's opinion as in favor or against a given target, which may or may not be present in the text [5].

Cotfas et al. [28] worked with tweets from the month following the Covid-19 vaccine announcement and found that the majority of tweets were in "neutral" territory and tweets in "favor" outpass "against" stance towards the vaccine.

In our project, we want to visually represent the change of people's stance on the topic of Covid-19 vaccines and show how the spread of misinformation topics over time impacts the opinion.

2.1 Datasets

We have prepared two separate datasets called the "primary dataset" and the "misinformation dataset" containing tweets during the Covid-19 pandemic. The datasets contain tweets, the date of the tweet, author locations in CSV format. Details of the two datasets are explained below.

Primary dataset

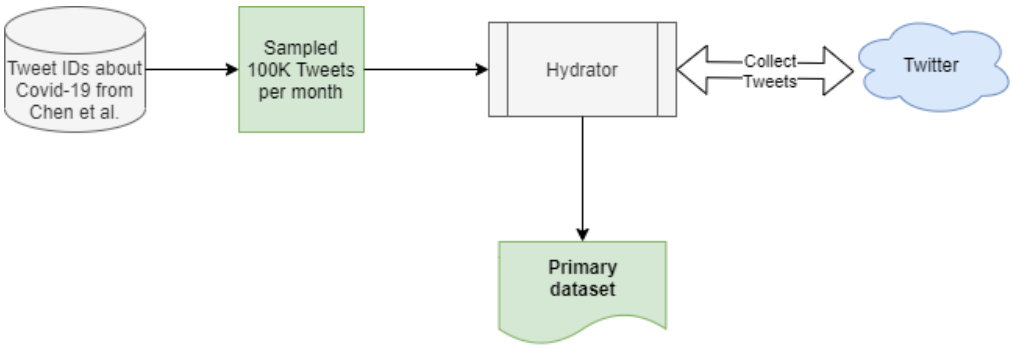

Figure 1: Steps to prepare the primary dataset

We have collected 6.2 million tweets between January 2020 and March 2021 that contain information about Covid-19 (SARS-COV-2). We have used the data from Chen et al. [12] and gathered the Tweet IDs from their GitHub repository [11]. Chen et al. used several keywords to search for the Covid-19 related tweets. Then we randomly sampled IDs from each month to fetch the detailed tweets. Finally used the Hydrator API tool [15] to collect all the tweet information. This is the primary dataset for the study.
Due to the rate limit of Twitter API [33], we sampled 100,000 tweets from each week and collected the complete tweet data using the above process. At the end of the process, we had 6,224,762 tweets in our dataset. Figure 1 shows the steps to prepare the primary dataset.

Misinformation dataset

To identify the prominent topics in misinformation, we first need to create a dataset containing the misinformation tweets relating to Covid-19. To achieve this, we collected the list of websites curated by NewsGuard [32] that are spreading Covid-19 related

misinformation. NewsGuard scores each website on a scale of 100 in 9 different metrics, as listed in Table 1 below.

Category	Score
Does not repeatedly publish false content	22
Gathers and presents information responsibly	18
Regularly corrects or clarifies errors	12.5
Handles the difference between news and opinion responsibly	12.5
Avoids deceptive headlines	10
Website discloses ownership and financing	7.5
Clearly labels advertising	7.5
Reveals who's in charge, including any possible conflicts of interest	5
The site provides names of content creators, along with either contact or biographical information	5
Total	100

Table 1: Metrics used by NewsGuard to score misinforming websites.

Tweets that contain links to these potential misinformation websites possibly contain misinformation content. We collected random tweet IDs from the year 2020 that contain any of the misinformation URLs, using the Brandwatch API [31], a web-based tool to collect historical social media data, to create the misinformation dataset. We used the Hydrator API tool to get the tweets from the tweet IDs. At the end of the process, we had 509,460 tweets in the misinformation dataset. Figure 2 shows the steps of preparing the misinformation dataset.

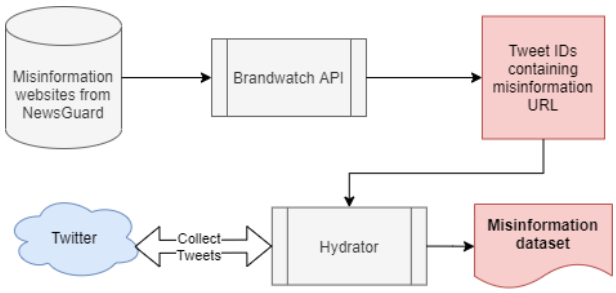

Figure 2: Steps to prepare the misinformation dataset

3 Project idea

3.1 Topic dataset

We will isolate the most prominent topics from each of the datasets in the data processing step. Then we can find the common topics from both datasets to identify the misinformation topics spreading in the Twittersphere and create a topic dataset. Figure 3 below shows the steps of the dataset creation.

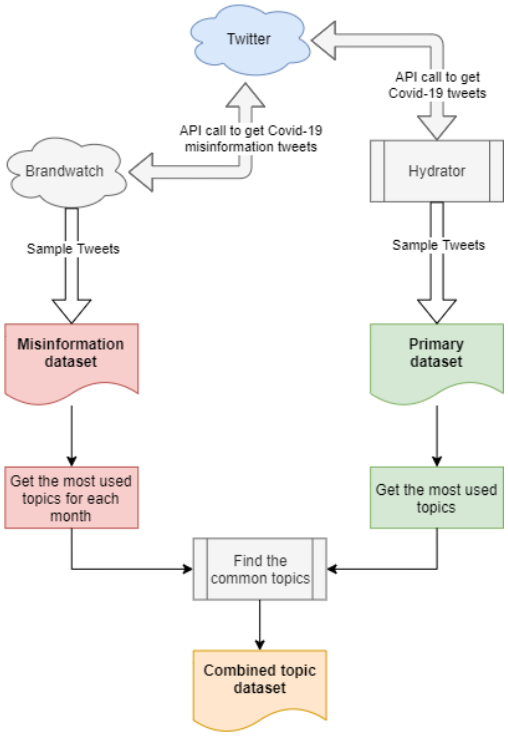

Figure 3: Steps to prepare the topic dataset

This dataset can give us a visual representation of the spread of misinformation over time. Figure 4 below shows a possible way to represent the spread of topics per month.

3.2 Sentiment analysis

From the tweets of the primary dataset, we isolated the English tweet texts and processed the texts through the VADER (Valence Aware Dictionary and sEntiment Reasoner) [3] sentiment analyzer. We chose VADER because it works well with social media text

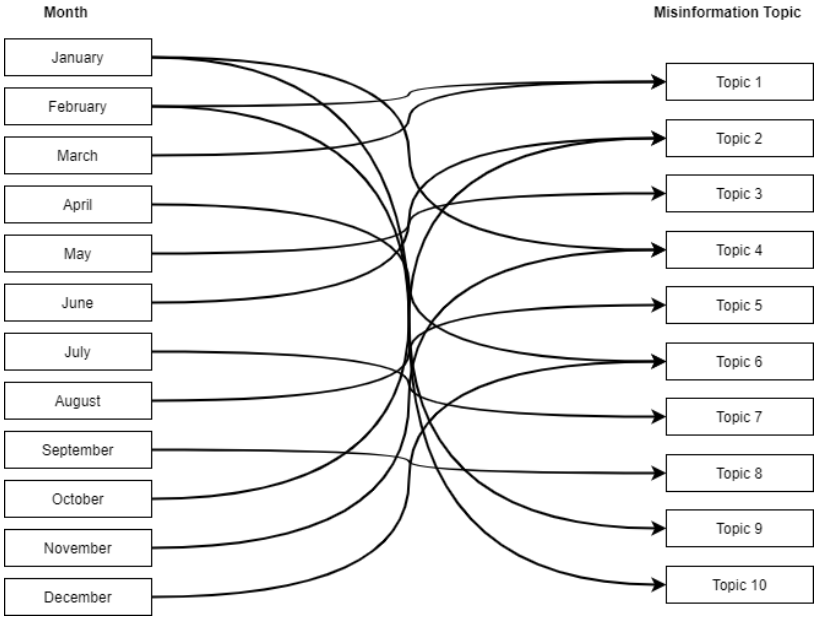

Figure 4: Possible way to visualize the misinformation topics per month

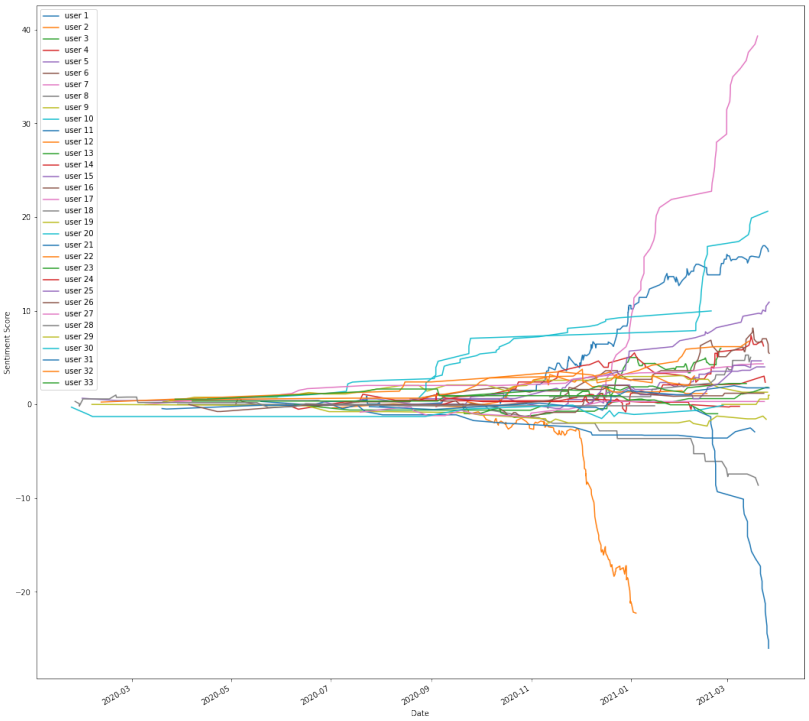

Figure 5: Cumulative sentiment score for Twitter users over time (user handles are anonymized per Twitter API guideline)

by considering different social media norms. This process gave us sentiment scores for each of the tweets. The Score data contained four fields - “pos,” “neg,” “neu,” and “compound.” The “pos,” “neg,” and “neu” represent the ratio of positive, negative, and neutral parts of the texts. The sum of these three fields becomes 1. The “compound” score is the most meaningful metric to measure the sentiment of the tweets. We chose the users who have tweeted 20 or more times about the vaccines and visualized a line chart of the cumulative “compound” score over time. Figure 5 shows that while most of the users were primarily neutral initially, they started to move towards the extremes of the positive and negative spectrum over time.

3.3 Goal

For our project, we propose the following goals.

1. Find suitable visualization methods to combine the spread of misinformation topics and the change of sentiment on Twitter.
2. If we can come up with multiple visualization ideas, run a user study to figure out the best method to convey the information

4 Timeline

The timeline we want to follow is explained in the table 2 below.

Date	Tasks
September 30, 2021	- Study previous works done in this area. - Find a novel approach or extension of an existing visualization method.
October 15, 2021	- Implement stance detection algorithms on the primary dataset. - Finalize the concept of the possible visualization methods to use.
October 30, 2021	- Develop the first draft of the visualization.
November 15, 2021	- Finalize the visualization. - Run user study if necessary.
November 30, 2021	- Finalize the project report and presentation.

Table 2: Timeline to complete the project tasks

References

[1] J. H. Kuklinski, P. J. Quirk, J. Jerit, D. Schwieder, and R. F. Rich, "Misinformation and the currency of democratic citizenship," *J. Polit.*, vol. 62, no. 3, pp. 790–816, 2000.

[2] L. Palen, *Online social media in crisis events*, https://er.educause.edu/articles/2008/8/online-social-media-in-crisis-events, Accessed: 2020-8-8, Aug. 2008.

[3] C. Hutto and E. Gilbert, "VADER: A parsimonious Rule-Based model for sentiment analysis of social media text," en, *ICWSM*, vol. 8, no. 1, pp. 216–225, May 2014.

[4] E. Housholder and H. L. LaMarre, "Political social media engagement: Comparing campaign goals with voter behavior," *Public Relat. Rev.*, vol. 41, no. 1, pp. 138–140, Mar. 2015.

[5] S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry, "SemEval-2016 task 6: Detecting stance in tweets," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 31–41. DOI: 10.18653/v1/S16-1003. [Online]. Available: https://aclanthology.org/S16-1003.

[6] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *SIGKDD Explor. Newsl.*, vol. 19, no. 1, pp. 22–36, Sep. 2017.

[7] A. Badawy, E. Ferrara, and K. Lerman, "Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Aug. 2018, pp. 258–265.

[8] D. A. Broniatowski, A. M. Jamison, S. Qi, L. AlKulaib, T. Chen, A. Benton, S. C. Quinn, and M. Dredze, "Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate," en, *Am. J. Public Health*, vol. 108, no. 10, pp. 1378–1384, Oct. 2018.

[9] V. L. Kitzie, E. Mohammadi, and A. Karami, "'life never matters in the DEMOCRATS MIND": Examining strategies of retweeted social bots during a mass shooting event," *Proc. Assoc. Info. Sci. Tech.*, vol. 55, no. 1, pp. 254–263, Jan. 2018.

[10] BBC News, "How do pandemics end?" *BBC*, Oct. 2020.

[11] E. Chen, *COVID-19-TweetIDs*, https://github.com/echen102/COVID-19-TweetIDs, Accessed: 2021-9-9, May 2020.

[12] E. Chen, K. Lerman, and E. Ferrara, "Tracking social media discourse about the COVID-19 pandemic: Development of a public coronavirus twitter data set," en, *JMIR Public Health Surveill*, vol. 6, no. 2, e19273, May 2020.

[13] W. Cornwall, "Officials gird for a war on vaccine misinformation," en, *Science*, vol. 369, no. 6499, pp. 14–15, Jul. 2020.

[14] A. Depoux, S. Martin, E. Karafillakis, R. Preet, A. Wilder-Smith, and H. Larson, "The pandemic of social media panic travels faster than the COVID-19 outbreak," en, *J. Travel Med.*, vol. 27, no. 3, May 2020.

[15] Documenting the Now, *Hydrator API*, 2020.

[16] E. Ferrara, "What types of COVID-19 conspiracies are populated by twitter bots?" en, *First Monday*, May 2020.

[17] A. M. Jamison, D. A. Broniatowski, M. Dredze, A. Sangraula, M. C. Smith, and S. C. Quinn, "Not just conspiracy theories: Vaccine opponents and proponents add to the COVID-19 'infodemic' on twitter," *HKS Misinfo Review*, Sep. 2020.

[18] S. Loomba, A. de Figueiredo, S. J. Piatek, K. de Graaf, and H. J. Larson, "Measuring the impact of exposure to COVID-19 vaccine misinformation on vaccine intent in the UK and US," Oct. 2020.

[19] G. Pennycook, J. McPhetres, Y. Zhang, J. G. Lu, and D. G. Rand, "Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable Accuracy-Nudge intervention," en, *Psychol. Sci.*, vol. 31, no. 7, pp. 770–780, Jul. 2020.

[20] C. A. Peschken, "Possible consequences of a shortage of hydroxychloroquine for patients with systemic lupus erythematosus amid the COVID-19 pandemic," en, *J. Rheumatol.*, vol. 47, no. 6, pp. 787–790, Jun. 2020.

[21] M. S. Al-Rakhami and A. M. Al-Amri, "Lies kill, facts save: Detecting COVID-19 misinformation in twitter," *IEEE Access*, vol. 8, pp. 155 961–155 970, 2020.

[22] S. Tasnim, M. M. Hossain, and H. Mazumder, "Impact of rumors or misinformation on coronavirus disease (COVID-19) in social media," Mar. 2020.

[23] E. K. Vraga and L. Bode, "Defining misinformation and understanding its bounded nature: Using expertise and evidence for describing misinformation," *Political Communication*, vol. 37, no. 1, pp. 136–144, Jan. 2020.

[24] WHO, *Coronavirus disease (COVID-19): Herd immunity, lockdowns and COVID-19*, https://www.who.int/news-room/q-a-detail/herd-immunity-lockdowns-and-covid-19, Accessed: 2021-3-4, Dec. 2020.

[25] X. Ye, B. Zhao, T. H. Nguyen, and S. Wang, "Social media and social awareness," in *Manual of Digital Earth*, H. Guo, M. F. Goodchild, and A. Annoni, Eds., Singapore: Springer Singapore, 2020, pp. 425–440.

[26] C. Aschwanden, "Five reasons why COVID herd immunity is probably impossible," en, *Nature*, vol. 591, no. 7851, pp. 520–522, Mar. 2021.

[27] CDC, *How to protect yourself & others*, https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/prevention.html, Accessed: 2021-9-16, Sep. 2021.

[28] L.-A. Cotfas, C. Delcea, I. Roxin, C. Ioanăş, D. S. Gherai, and F. Tajariol, "The longest month: Analyzing COVID-19 vaccination opinions dynamics from tweets in the month following the first vaccine announcement," *IEEE Access*, vol. 9, pp. 33 203–33 223, 2021.

[29] C. Wardle and E. Singerman, "Too little, too late: Social media companies' failure to tackle vaccine misinformation poses a real threat," en, *BMJ*, vol. 372, n26, Jan. 2021.

[30] WHO, *Coronavirus disease (COVID-19): Vaccines*, https://www.who.int/news-room/q-a-detail/coronavirus-disease-(covid-19)-vaccines, Accessed: 2020-10-28, Jun. 2021.

[31] Brandwatch, *Brandwatch: A new kind of intelligence*, https://www.brandwatch.com/, Accessed: 2021-1-27.

[32] NewsGuard, *Coronavirus misinformation tracking center - NewsGuard*, https://www.newsguardtech.com/coronavirus-misinformation-tracking-center/, Accessed: 2021-1-27.

[33] Twitter, *Twitter developer*, https://developer.twitter.com/, Accessed: 2020-1-25.