Literature Review

Implementing Cyber Security tools and/or techniques in detecting medical misinformation on a social media platform of Twitter

Ying Chan

Hong Kong

University of Essex

Research Methods and Professional Practice November 2022

Unit 7

Table of Contents

[Abstract 3](#_Toc124784345)

[Introduction 3](#_Toc124784346)

[Prevalence of Medical Misinformation on a Social Media Platform 3](#_Toc124784347)

[The Medical Misinformation that Propagation for COVID-19 4](#_Toc124784348)

[Effects of medical misinformation 5](#_Toc124784349)

[Health and Life Matters 5](#_Toc124784350)

[Social Awareness 5](#_Toc124784351)

[Vaccines 5](#_Toc124784352)

[Ways to ban or prevent such misinformation on Twitter 6](#_Toc124784353)

[Post Labelling 6](#_Toc124784354)

[Violation on Content 6](#_Toc124784355)

[Accounts Control 6](#_Toc124784356)

[Comparison of other social media on medical misinformation 7](#_Toc124784357)

[World Health Organization 7](#_Toc124784358)

[Facebook 7](#_Toc124784359)

[YouTube 7](#_Toc124784360)

[Techniques of detecting medical misinformation 8](#_Toc124784361)

[Grading Post 8](#_Toc124784362)

[Artificial Intelligence (AI) 9](#_Toc124784363)

[Machine learning (ML) 9](#_Toc124784364)

[Possible Area for Utilising Cyber Security Tools 10](#_Toc124784365)

[Manipulated Audio and Video 10](#_Toc124784366)

[Blockchain Technology 10](#_Toc124784367)

[Conclusion 10](#_Toc124784368)

[References 11](#_Toc124784369)

# Abstract

The prevalence, relationships, and effects of medical misinformation are examined in this research. Learned how to use Cyber Security technologies to find medically inaccurate tweets on the social media site Twitter. Review the methods and compare the preventative measures to those used on other social platforms. Implementing Cyber Security tools revealed the improvement.

# Introduction

## Prevalence of Medical Misinformation on a Social Media Platform

Gunther's investigation indicates the importance of social media sites like Twitter, Facebook, and YouTube for the rapid and extensive dissemination of information. A rise in improperly understanding scientific knowledge, the polarization of ideas, an increase in dread and panic, or a decrease in access to healthcare are only a few effects of sharing false information on social media. Gunther's systematic research projects provide a thorough analysis of the most common health myths as well as a breakdown of how prevalent they are across various social media channels. The study found that topics including medicines and smoking goods had the highest frequency of health misinformation on Twitter. However, there was also a lot of disinformation out there about important public health concerns, such as infections and immunizations (Gunther, 2021).

## The Medical Misinformation that Propagation for COVID-19

As of 13 January 2023, there have been 661 million confirmed cases of COVID-19, including 6 million fatalities, from 194 Member States since the outbreak of COVID-19 began in December 2019 (WHO, 2023).

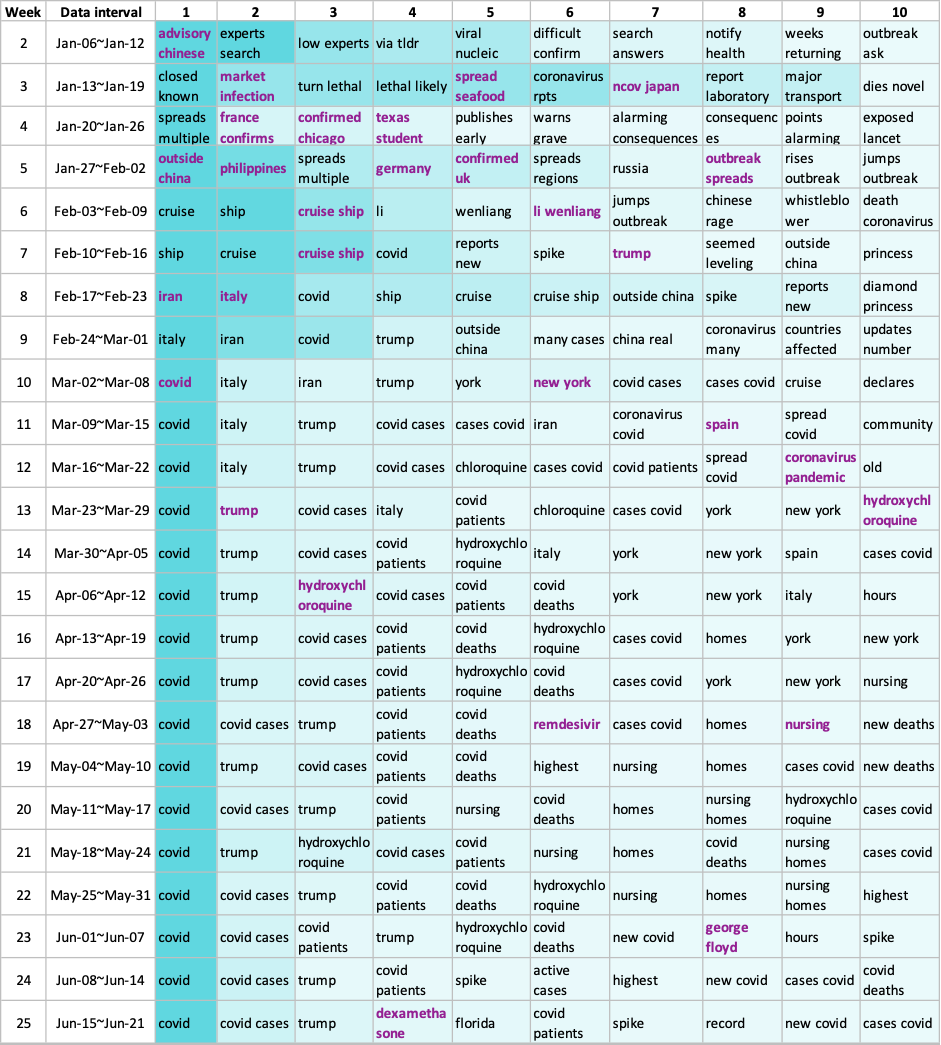


Table 1: Top 10 topics of each week (Rita & Corey, 2021).

The top 10 topics in the term frequency-inverse document frequency (tf-idf) scores, as reported by the Journal of Medical Internet Research, are displayed in Table 1 (Rita & Corey, 2021). Without a doubt, COVID-19 subjects have attracted the most medical content on social media in recent years.

# Effects of medical misinformation

## Health and Life Matters

The spread of false information about health has the potential to kill both direct and indirect. False perceptions of the threat, unsuccessful methods of coping, and a variety of catastrophic outcomes can result from exposure to misleading information about the COVID-19 pandemic's spreading, infection symptoms, testing probability, and prophylactic actions. The public may become suspicious of vaccines due to incorrect information regarding the new COVID-19 vaccines and the ongoing development, which is more significant and could hinder vaccination rates that are significant enough to give herd immunity (Sun, et al., 2022).

## Social Awareness

Due to the significant amount of uncertainty regarding the pandemic and the relatively fast pace of vaccine development compared to other types of regular vaccines, the public naturally sought out information to allay their vaccination concerns and help guide important decisions like whether to get vaccinated or not. With so much information being disseminated through media organizations and webpages with varying degrees of validity and truth, it can be difficult to distinguish between real and accurate information about COVID-19 immunizations and misleading and inaccurate falsehoods. A main challenge to finding accurate, reliable information on the COVID-19 vaccine is the pervasive, unsolicited, and dubious pseudo-news articles posted on social media and other internet platforms by various kinds of individuals (Sun, et al., 2022).

## Vaccines

The COVID-19 vaccine misinformation that is spread among the American population was explored in the scientific report's study, along with how accuracy in the understanding the COVID-19 vaccinations relates to vaccine reluctance and behavioural intention. The knowledge test used in the quantitative method also revealed that many people accepted false information about the COVID-19 vaccine as fact, and that faulty information tended to raise people's vaccine reluctance and lower their behavioural intention to be immunized. The results of the study corroborated other studies' conclusions about the significance of correct information and the impact of false information on vaccination hesitation and refusal (Sun, et al., 2022).

# Ways to ban or prevent such misinformation on Twitter

## Post Labelling

When content is not removed that violates Twitter's policy, Twitter may give further information to that misleading media when those Tweets appear on the platform (Twitter, 2022).

The series of actions comprised:

* Label and/or warn the Tweet.
* Warning people who would like to share or like the Tweet.
* Make the Tweet less visible on Twitter and/or stop people from recommending it.
* Disable retweets, replies, and likes.
* Include a link to more information or clarifications.

## Violation on Content

Profile: If profile material such as a bio violates Twitter's policy, Twitter will request that the content be removed. Before sending the next tweet, Twitter will also temporarily freeze the account. After receiving the first warning, the account will be permanently suspended if you breach this policy once more.

Post: Twitter will be required to remove this content for high-severity policy breaches, such as false media that poses a severe risk of harm to individuals or communities (Twitter, 2022).

## Accounts Control

Qualification: Twitter targets to lessen harm by restricting the number of times offensive information is displayed on the site. Tweets from high-profile accounts, such as verified, official government accounts, state-affiliated media accounts, or Tweets with many impressions or interactions, will be prioritized by Twitter for enforcement.

Strike system: When accounts regularly break Twitter's policy, Twitter will utilize a strike system to decide whether more enforcement measures need to be taken. A 12-hour account time-out will be imposed when two notices are applied within 30 days. Seven-day time-outs will be imposed if three or more notifications are issued within 30 days. A permanent suspension will be imposed after five strikes or more (Twitter, 2021).

# Comparison of other social media on medical misinformation

## World Health Organization

The World Health Organization (WHO) and its partners are aware that false information on the internet has the potential to spread further, quicker, and occasionally deeper than the truth. For example, on Facebook and YouTube, misleading information is 70% more likely to be shared than reliable news. To combat this, WHO has collaborated with digital businesses on several initiatives. (WHO, 2023).

## Facebook

In collaboration with national and local health authorities, Facebook identifies public health crises. This includes false statements about COVID-19 that have been confirmed by reputable medical authorities, including those regarding the virus' existence or severity, its treatment or prevention, its transmission or immunity.

When public health authorities determine that false information is likely to directly increase the risk of impending physical harm, including by increasing the likelihood that people will contract or spread a harmful disease or reject an associated vaccine during a public health emergency, Facebook removes the misinformation (Facebook, N.D.).

## YouTube

COVID-19 Medical Misinformation Policy: Forbids the publication of anything that disproves the medical advice provided by national and international health authorities about COVID-19.

Vaccine Misinformation Policy: Prohibits the dissemination of information that poses a serious threat of severe harm by spreading erroneous information about currently administered vaccines that are recognized safe and effective by national and international health authorities (WHO). This only applies to information that contradicts the safety, efficacy, and composition recommendations for vaccinations made by the WHO or local health authorities (YouTube, N.D.).

# Techniques of detecting medical misinformation

## Grading Post

In contrast, Twitter lifts its restriction on false material about COVID-19. Appears to be delegating greater control over policing false user data through the company's Birdwatch initiative, which enables Twitter users to grade and contribute fixes to tweets (Taylor, 2022). The outcome is reflected in the Post Labelling.

Over 100 qualitative interviews with people from all political backgrounds were done by Twitter in 2021 and the Birdwatch garnered strong approval across the board. People enjoyed that comments gave helpful background to assist people in better comprehending and assessing a Tweet rather than focusing on labelling information as true or false. Those notes were written from the community's perspective rather than that of Twitter or central authority. The intention of Birdwatch is for the Twitter community to shape it in an open environment (Twitter, 2021).

The findings of four studies conducted on various dates between 2021 and 2022 show that seeing a Birdwatch note reduces one's likelihood of agreeing with the message of a possibly deceptive Tweet by, on average, 20 to 40%. Between 3,000 and over 19,000 people participated in the survey, and even if news and tweet themes changed throughout the course of the past year, the results remained steady (Twitter, 2022).

In Twitter policy, users must first demonstrate their note-writing skills to contribute to Birdwatch. Twitter determines this by allocating a "rating impact" score to each potential contributor. By ranking Birdwatch remarks as "Helpful" or "Not Helpful," users can earn or lose points (Twitter, 2022). Despite this, some individuals could question if false material can be added to tweets only because many abused users gave them a high rating.

## Artificial Intelligence (AI)

When it comes to spotting fake news on Twitter during the Covid-19 era, Madani, Youness demonstrates a novel method. The experimental findings indicate that when combined with the random forest method, it yields extremely beneficial results with an accuracy of 79%. Additionally, they show how essential sentiment is in identifying bogus news in tweets. In fact, the model they describe performs better than existing models since it considers the features of tweets (Youness, Mohammed & Belaid, 2021).

## Machine learning (ML)

In Sreedevi Marriboyina's study, they discovered that one of their machine learning approaches, Support Vector Machine (SVM), had roughly 99.48% accuracy when they just used Twitter data for the prediction of false tweets. It is a method of supervised learning that may be applied to relapse and order problems. The datasets are shown as points. The primary goal of SVM is to create a hyperplane that categorizes the datasets into separate groups. The hyperplane, a flexible boundary that aids in grouping informational foci, should be placed at the boundary from classes. This approach improves accuracy by discarding functions that are too tightly fitted to a small number of sampled data points (Sreedevi, 2021).

# Possible Area for Utilising Cyber Security Tools

## Manipulated Audio and Video

Besides text content, audio and video information also spread quickly because it catches people's attention and is shared often. The use of manipulated audio and video footage to disseminate misleading information is extremely harmful.

Cheap fakes: Real audio and video recordings that have been accelerated, slowed down or exhibited out of context to deceive.

Deepfakes: Audio and video recordings that are manufactured to be deceptive yet sound extremely real. They have the power to persuade that others have spoken or done things that never occurred. Visual Deepfakes may produce full-body videos or fake with convincing faces. A voice replica that creates fresh phrases from one or more persons on its own or in conjunction with a false video is known as an audio Deepfake (CISA, N.D.).

## Blockchain Technology

In recent research, a distributed data-sharing system based on smart contracts is suggested to counteract Deepfakes. By authenticating content, blockchain technologies offer tamper-proof assurance that it is legitimate. A smart contract that is used to make videos stores the video's information and any associated properties before producing a hash. Additionally, the smart contract provides secure access to the movie for a certain group of people. Because each user who wishes to download or view the film must engage with the smart contract, the authenticity of such video is preserved and Deepfakes are prevented (Abbas, et al., 2021).

# Conclusion

Machine learning or artificial intelligence is one of the defences used to present suspicious posts for voting. Use an AI/ML-based system that investigates the characteristics of tweets and scans the content for keywords. By comparing the phoney material with authentic or legitimate content, this process may be utilized to find abnormalities and notify the appropriate authorities. However, there is a restriction on video analysing that must convert to text or image.

Other than using AI/ML technologies to determine if text content has been altered. Blockchain technology, one of the cyber security tools, maybe a clever approach to demonstrate the authenticity of material, particularly video. To uphold the principles of openness and transparency, anomalies should be processed to the next stage for human intervention with Twitter's Birdwatch rating system serving as the final line of defence.

# References

Abbas, Y. et al. (2021) Making Sense of Blockchain for AI Deepfakes Technology. *2020 IEEE Globecom Workshops*. DOI: https://doi.org/10.1109/GCWkshps50303.2020.9367545

CISA. (N.D.) Inauthentic Content. *Tools of Disinformation*. Available from: https://www.cisa.gov/sites/default/files/publications/mdm-inauthentic-content-product-english\_508.pdf [Accessed 13 January 2023].

Facebook. (N.D.) Misinformation. *Facebook Community Standards*. Available from: https://transparency.fb.com/en-gb/policies/community-standards/misinformation/ [Accessed 13 January 2023].

Gunther, E. (2021) Prevalence of Health Misinformation on Social Media: Systematic Review. *J Med Internet Res*. 23(1): e17187. DOI: https://doi.org/10.2196%2F17187

Rita, K. & Corey, B. (2021) Using Tweets to Understand How COVID-19–Related Health Beliefs Are Affected in the Age of Social Media: Twitter Data Analysis Study. *J Med Internet Res*. 23(2): e26302. DOI: https://doi.org/10.2196%2F26302

Sreedevi, M. (2021) Prediction of Fake Tweets Using Machine Learning Algorithms. *Recent Trends in Intensive Computing*. DOI: http://dx.doi.org/10.3233/APC210195

Sun, L. et al. (August 11, 2022) Misinformation of COVID-19 vaccines and vaccine hesitancy. *Scientific Reports* 12, 13681. DOI: https://doi.org/10.1038/s41598-022-17430-6

Taylor, L. (November 29, 2022) Twitter ends its ban on covid misinformation. *Technology*. Available from: https://www.washingtonpost.com/technology/2022/11/29/twitter-covid-misinformation-policy/ [Accessed 13 January 2023].

Twitter. (2022) Synthetic and manipulated media policy. *Rules and Policies*. Available from: https://help.twitter.com/en/rules-and-policies/manipulated-media [Accessed 13 January 2023].

Twitter. (January 25, 2021) Introducing Birdwatch, a community-based approach to misinformation. *Product*. Available from: https://blog.twitter.com/en\_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation [Accessed 13 January 2023].

Twitter. (October 6, 2022) Helpful Birdwatch notes are now visible to everyone on Twitter in the US. *Product*. Available from: https://blog.twitter.com/en\_us/topics/product/2022/helpful-birdwatch-notes-now-visible-everyone-twitter-us [Accessed 13 January 2023].

Twitter. (October, 2021) Civic integrity policy. *Rules and Policies*. Available from: https://help.twitter.com/en/rules-and-policies/election-integrity-policy [Accessed 13 January 2023].

Twitter. (September 7, 2022) Birdwatch is getting a new onboarding process and more visible notes. *Product*. Available from: https://blog.twitter.com/en\_us/topics/product/2022/birdwatch-getting-new-onboarding-process-more-visible-notes [Accessed 13 January 2023].

WHO. (January 13, 2023) Global Situation. *WHO Coronavirus (COVID-19) Dashboard*. Available from: https://covid19.who.int/ [Accessed 13 January 2023].

Youness, M., Mohammed, E. & Belaid, B. (2021) Using artificial intelligence techniques for detecting Covid-19 epidemic fake news in Moroccan tweets. *Results in Physics* Volume 25 104266. DOI: https://doi.org/10.1016/j.rinp.2021.104266

YouTube. (N.D.) How does YouTube address misinformation?. *Our Commitments*. Available from: https://www.youtube.com/howyoutubeworks/our-commitments/fighting-misinformation/#misinformation-policies [Accessed 13 January 2023].