

**A FOREST-FIRE MODEL APPROACH ON
UNDERSTANDING THE DYNAMICS OF
INFORMATION SPREAD ON SOCIAL MEDIA**

PATRICK OWEN KUEK

This project is submitted in partial fulfilment of the
requirements for the degree of
Bachelor of Computer Science with Honours
(Computational Science)

Faculty of Computer Science and Information Technology
UNIVERSITI MALAYSIA SARAWAK
2025

DECLARATION

I hereby declare that this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that no portion of the work referred to in this report has been submitted in support of an application for another degree at Universiti Malaysia Sarawak (UNIMAS).

Signed,

A handwritten signature in black ink, appearing to be 'A' with a stylized flourish.

(Patrick Owen Kuek, 80944)

21st July 2025

Faulty of Computer Science and Information Technology

Univeristi Malaysia Sarawak

ACKNOWLEDGEMENT

I would like to express my appreciation to all who have contributed towards the success of this final year project. Firstly, I would like to acknowledge my deepest thanks to my family for their huge support not only throughout the project, but also throughout the entire four academic years. Their belief in my determination has been a source of my strength to keep pushing forward during the hard times of my academic life.

Next, I am especially thankful to my supervisor, Professor Dr. Jane Labadin, for her invaluable guidance and expertise. Her feedback and constructive criticism throughout the process have improved the quality of my work. I appreciate the time she dedicated to reviewing my progress and the patience she showed as I explained my concept that might have been confusing.

Last but not least, I would like to extend my gratitude to my friends for their encouragement and companionship through thick and thin. Whether it was staying up at night to complete assignments or shared laughter, their support has made my academic life very enjoyable and fulfilling.

ABSTRACT

The usage of social media has become increasingly prevalent, leading to a continuous rise in information sharing which can have both positive and negative impacts on society. In this project a forest-fire model is developed to study the dynamics of the information spread on social media. The factors that influence the spread are the user activity (UA), follower-following ratio (FF), and topic significance (TS). These factors are combined using weighted parameters to determine the probability of information spreading between users. Results show that the model exhibits accuracy when the network's time window connection is synchronous with the time segment of the observed dataset. Additionally, Sobol sensitivity analysis reveals that the spread of the model is heavily influenced by the weight corresponding to the topic significance (TS). The findings provide insights into the behavioural dynamics of online information spread and highlight the importance of content relevance in social media.

ABSTRAK

Peningkatan dalam penggunaan media sosial telah membawa kepada peningkatan berterusan dalam perkongsian maklumat yang boleh membawa pelbagai kesan positif dan negatif kepada masyarakat. Untuk projek ini, model kebakaran hutan digunakan untuk mengkaji dinamik trend penyebaran maklumat dalam media sosial. Antara faktor yang mempengaruhi penyebaran maklumat ialah aktiviti pengguna media sosial (UA), nisbah pengikut-kepengikutan pengguna (FF), dan kepentingan topik maklumat (TS). Ketiga-tiga faktor ini akan digunakan dalam pemodelan kebarangkalian maklumat tersebar antara pengguna dengan menggunakan pemberat sebagai parameter model. Hasil projek menunjukkan bahawa model tersebut memberikan ketepatan yang tinggi apabila tingkap masa yang digunakan dalam perangkaan adalah selari dengan pecahan masa dalam set data pemerhatian. Tambahan pula, analisis sensitiviti Sobol menunjukkan bahawa penyebaran maklumat dalam model ini amat dipengaruhi oleh pemberat yang mewakili kepentingan topik maklumat (TS). Dapatan daripada projek ini mencetus pelbagai pandangan terhadap dinamik penyebaran maklumat dalam talian serta menekankan kepentingan kerelevanan topik dalam media sosial.

TABLE OF CONTENTS

	Page
DECLARATION	i
ACKNOWLEDGEMENT	ii
ABSTRACT	iii
<i>ABSTRAK</i>	iv
TABLE OF CONTENTS	v
LIST OF TABLES	vii
LIST OF FIGURES	viii
LIST OF ABBREVIATIONS	ix
CHAPTER 1: INTRODUCTION	1
1.1 Background	1
1.2 Problem Statement	3
1.3 Objectives	4
1.4 Scope	4
1.5 Brief methodology	4
1.6 Significance of project	7
1.7 Project schedule	7
1.8 Expected outcome	7
CHAPTER 2: LITERATURE REVIEW	8
2.1 Introduction	8
2.2 Epidemic model	8
2.2.1 Classic epidemic model	9
2.2.2 Improved epidemic model	10
2.3 Heat energy diffusion model	12
2.4 Network model	13
2.4.1 Forest fire model	13
2.5 Summary	14
CHAPTER 3: METHODOLOGY	16
3.1 Introduction	16

3.2	Dynamics of information spreading on social media	18
3.3	Modelling	20
3.3.1	Defining of spread probability, p	20
3.3.1.1	User activity (UA)	20
3.3.1.2	Topic significance (TS)	21
3.3.1.3	Follower-following ratio (FF)	22
3.3.2	Data collection	24
3.3.2.1	Dataset description	24
3.3.2.2	Data preprocessing	26
3.3.3	Forest fire model	28
3.4	Summary	29
CHAPTER 4: IMPLEMENTATION & ANALYSIS		30
4.1	Introduction	30
4.2	Environment setup	30
4.3	Forest-fire algorithm	31
4.4	Sensitivity analysis	34
4.5	Summary	35
CHAPTER 5: RESULTS & DISCUSSION		36
5.1	Introduction	36
5.2	Model results	36
5.3	Sensitivity analysis results	41
5.4	Summary	42
CHAPTER 6: CONCLUSION & FUTURE WORKS		44
6.1	Introduction	44
6.2	Achievement of objectives	44
6.3	Project limitations	45
6.4	Future works	46
6.5	Summary	47
REFERENCES		48
APPENDICES		52

LIST OF TABLES

		Page
Table 3.1	Parallels between forest fire spread and information spread	19
Table 3.2	Ranking of scores based on topic significance	21
Table 3.3	Range of TS values according to ranking	22
Table 3.4	Parameter definition and its ranges	24
Table 5.1	MAE of each model result	39
Table 5.2	Sobol results under different configurations	43
Table 6.1	Objectives and its corresponding achievements	44

LIST OF FIGURES

		Page
Figure 1.1	Modelling and simulation scheme	5
Figure 2.1	Diagram of SI, SIS, and SIR model respectively	9
Figure 3.1	Flowchart of the methodological process of the project	17
Figure 3.2	Information spread in social network	18
Figure 3.3	Example posts from X regarding the <i>Jalur Gemilang</i> badge policy	25
Figure 3.4	Flowchart of cleaning the raw dataset	26
Figure 4.1	Forest-fire algorithm	32
Figure 5.1	Number of spreaders over eight-hour intervals	36
Figure 5.2	Number of spreaders over eight-hour intervals	37
Figure 5.3	Number of spreaders over eight-hour intervals	37
Figure 5.4	Model results observed under 8-hour segments	38
Figure 5.5	Model results observed under 6-hour segments	38
Figure 5.6	Model results observed under 12-hour segments	39
Figure 5.7	Output model with the most optimal weight combination	40

LIST OF ABBREVIATIONS

ARIMA	Auto-Regressive Integrated Moving Average
CPU	Central Processing Unit
E	Exposed
FF	Follower-Following ratio
GPU	Graphics Processing Unit
ISR	Ignorant-Spreader-Recovered
I	Infected
ID	Identification
MAE	Mean Average Error
R	Recovered
S	Susceptible
SAIR	Susceptible-Active-Infected-Recovered
SEIR	Susceptible-Exposed-Infected-Recovered
SEIRS	Susceptible-Exposed-Infected-Recovered-Susceptible
SEPNS	Susceptible-Exposed-Positive-Negative-Susceptible
SETQR	Susceptible-Infected-Trust-Question-Recovered
SI	Susceptible-Infected
SIDR	Susceptible-Infected-Doubt-Recovered
SIS	Susceptible-Infected-Susceptible
SIR	Susceptible-Infected-Recovered
SVEIR	Susceptible-Verified-Exposed-Infected-Recovered
TS	Topic Significance
UA	User Activity

CHAPTER 1: INTRODUCTION

1.1 Background

The evolution of communication is pivotal for human development, transforming how information is shared and understood by people across the globe. The rapid advancements in the communication domain have greatly changed the way information is spread. The methods of information dissemination have transitioned from traditional formats, like newspapers and radio, to modern digital platforms such as social media (Rosmani et al., 2020). Social media is a type of online mass media communication where people exchange ideas, content, personal messages, and other information. Unlike traditional media, where communication only happens one-way from sender to recipient without the latter's feedback, the two-way communication in social media allows for an immediate, interconnective, interactive, and user-driven information sharing experience for the people (Al-Quran, 2022). Nowadays, it has become an integral part of daily life, leading to an ever-rising trend in information sharing across platforms like X, Facebook, Instagram, and TikTok.

Information can be positive, such as ideas, news, and opinions, or negative, such as rumours and gossips. The shift to information spreading in social media, be it positive or negative, can greatly influence in how people engage with the information, hence affecting their lives (Wen et al., 2015). For example, it affects the way how businesses operate and interact with their customers. According to Gopal & Kovoov (2022), social media has been a powerful business tool, enabling companies to market their products, analyse consumer behaviour through data insights, and provide real-time customer support. However, with the growth of the business popularity, it could have the potential to be injected with rumours, such

as fake postings, in order to take down its competition. Another negative example is the spread of misinformation of the COVID-19 disease all over various social media platforms when the pandemic was at its peak, hindering the treatment of the public and cause high fatalities because of what they had heard from the internet (Ojha et al., 2023). As the spread of information across social media continues to grow, so does its spread of negative information. Because of this, it has posed challenges not only in businesses and public health, but also in politics and social cohesion, thus making it crucial to understand the dynamics of information spread to develop effective measures for mitigating the negative impacts.

Numerous studies have explored the dynamics of information dissemination on social media platforms using mathematical modelling to provide a better understanding of the spreading, to plan and control its progression (Sahnoune et al., 2021). The study on information dissemination was firstly presented by Daley and Kendall (1967) where they made attention to the familiarity between epidemics and rumours, thus creating a mathematical model similar to the epidemiological model by interpreting the population as ignorants, spreaders, and stiflers (Daley and Kendall, 1967). Modifications has been made decades after the original model was proposed, such as the inclusion of ‘verified’ population, who verifies information before sharing, where it is concluded that the chances to spread misinformation decreases as the number of ‘verified’ individuals increases (Ojha et al., 2023). There are other models developed apart from using the epidemiological approach. For example, a Markov chain model approach was used to generalise the growth with the inclusion of parameters of non-traditional behaviours like forgetfulness, apathy, and losing and regaining interest (de Arruda et al., 2018). Though countless models have been developed, it often lacks practical application due to the lack of validation. According to Wen et al. (2015), previous models have primarily focused on

simulations without real-world validation. This highlights the need for models that not only simulate scenarios but are also validated through real-world data to ensure practical applicability.

The project aims to develop a mathematical model that analyses the dynamics of information spreading on social media, primarily on X, focusing on user engagement, and to validate the model with actual X data to assess its accuracy. The developed model would enable individuals to predict the outcomes of information spread by simulating various scenarios and dynamics. This helps uncover how changes in one part of the system could affect the whole, thus offering valuable insights to make effective strategies for communication, education, and public engagement, particularly in combating misinformation.

1.2 Problem Statement

With the rapid development of communication technology over the years, information is rapidly spreading across various social media platforms. Since almost everyone now owns a smartphone, social media, without a doubt, has greatly influenced our lives when it comes to receiving information, whether it is positive or negative. The study of information spreading on online social media has attracted attention from various research fields because of its significant impact on the public and the challenges posed by its large volume and dynamic nature (Dey et al., 2017). Despite numerous studies, several existing models often lack adaptability to real-time social dynamics and varying network structures, limiting their practical application in managing the spread of information. By analysing the growth patterns, it could provide potential information that could benefit society that seeks to manage and mitigate the negative impacts of information spreading on social media.

1.3 Objectives

The objectives to be achieved in this project are as follows:

- i. To create a computational model that simulate the dynamics of information spreading on social media.
- ii. To validate the simulated model against real X data to assess its predictive accuracy.
- iii. To investigate the impact of each identified parameter on the information spread through sensitivity analysis.

1.4 Scope

The project will primarily focus on analysing X (previously known as Twitter) data collected over a specific time frame to investigate the user interaction and information spreading patterns. The project will develop and validate a mathematical model based on the observed spreading patterns, focusing solely on X data. Due to the vast amounts of social media data, the analysis will be limited to a specified time period and focuses on the spread of one type of information.

1.5 Brief Methodology

To achieve the objectives, the project follows a methodology based on the modelling and simulation scheme proposed by Velten (2009) as illustrated in Figure 1.1.

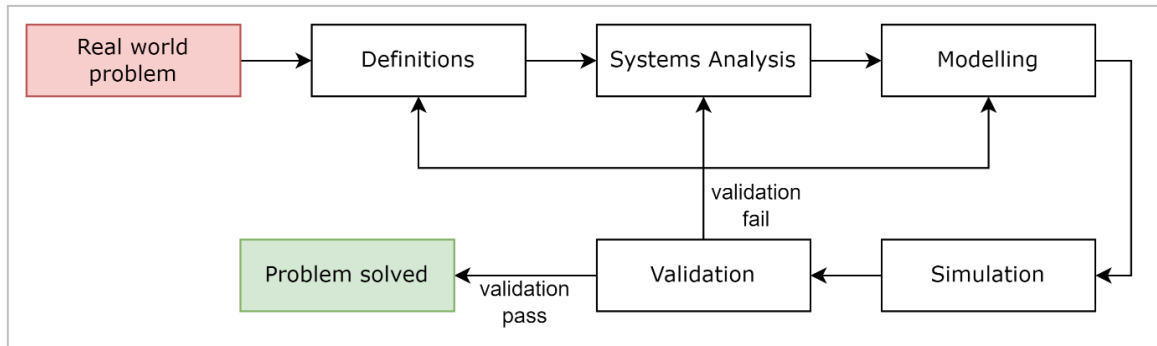


Figure 1.1: Modelling and simulation scheme

i. Definitions

In the beginning of the modelling phase, real world problems are identified represented as a system, so that it can be solved and analysed using mathematical models. By defining the problem, along with its goals and objectives, the project can establish a clear focus for the model, ensuring that it accurately represents the dynamics of the situation and addresses the specific aspects of the issue.

ii. Systems Analysis

In this phase, the focus is on identifying and understanding the key components, factors, and relationships of the system. This includes reviewing existing literature to learn from and apply approaches that have been used to address similar problems.

Additionally, data handling is done by scraping social media data and pre-processed it to ensure that the model is based on accurate and relevant data, which is essential for its credibility and relevance. In this project, X data, focusing on one specific type of information, is selected as the primary data

source. Data is collected within a range of at least two weeks to capture the historical trends. The collected dataset consists of posts containing a specified topic, or keyword, and additional metrics such as account username, user following count, user follower count, and the date created.

The data is then pre-processed, which include data cleaning to remove duplicate entries, such as posts created by the same account on the same day and time. Once the data is cleaned, it will then be used for modelling.

iii. Modelling

The modelling phase involves developing an agent-based social network model using on the data and information gathered from the previous step. A forest fire algorithm is applied in this project because the propagation of the information is more focused on individual interactions, thus reflecting a more accurate representation on how posts are spread between users. Therefore, the probability of spreading of information by a user has to be formulated.

iv. Simulation

In the simulation phase, the model will be simulated in a simulation environment, such as Python, to observe how the information spread under different scenarios, providing more insights on the dynamics of the system.

v. Validation

The predictive model will be compared with the pre-processed X data to assess its accuracy using statistical methods. With adequate accuracy, the model

is can be proved beneficial for future studies. However, if the validation fails with low accuracy, the model has to be refined by backtracking to previous steps such as improving the model formulation, reanalyse the system, or even redefine the problem. If the validation is successful, the model will then be used to perform sensitivity analysis to gain insights on how each contributing factor influences the information spread.

1.6 Significance of Project

By developing a robust model and validating it with real data, it has the potential to enhance understanding of information spread in social media. It could provide insights into controlling misinformation in its early stages and optimising public communication strategies to drive positive social behaviour. Additionally, it aims to fill a key research gap, contributing a validated approach to studies that have so far relied solely on simulations.

1.7 Project Schedule

The project schedule with Gantt chart is shown in Appendix A.

1.8 Expected Outcome

The outcome of this project is a well-structured model that represents the dynamics of information spreading. The model should be validated against real X data. To ensure the reliability of the model, the mean absolute error has to be low.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

Information dissemination, or information propagation, refers to the spread or distribution of information to other people. The rapid growth of social media has transformed how information is shared, consumed, and perceived. When information spreads widely, it can go viral or create a trend that quickly spreads, resembling a social epidemic (Yang et al., 2023). The dynamics of information dissemination on social media platforms have attracted significant attention from researchers across various disciplines (Dey et al., 2017). To better understand the underlying mechanisms governing this process, they have developed information dissemination models. Mathematical models play an important role in this effort, providing a structured framework to analyse this complex social phenomenon. This literature review examines and summarises various approaches researchers have taken to develop the model.

2.2 Epidemic Model

Epidemic models are mathematical frameworks designed to simulate the spread of infectious diseases within a population. These models can also be applied beyond disease outbreaks to scenarios where the concept of spreading is similar, such as the spreading of information. The study of information dissemination was firstly explored by Daley and Kendall (1967), who pointed out a similarity between the spread of infectious diseases and the spread of rumours. In their model, known as the DK model, they categorised the population into ignorants, spreaders, and stiflers, which corresponds to the susceptible, infected, and recovered

states, respectively, in future works. The dynamics of how users transitioned between these states are relatively simple to understand, making it a straightforward approach to studying the spread of information within a population. According to Liu et al. (2023), the epidemic model is widely used by many researchers in studying the information dissemination in social media.

2.2.1 Classic Epidemic Model

The most common epidemic model is the Susceptible-Infected-Recovered (SIR) model. Other classic models include the Susceptible-Infected (SI) model, and the Susceptible-Infected-Susceptible (SIS) model. Based on the classic models, individuals typically start in the susceptible state, where they are unaware of the information. They become infected, begin actively spreading the information, through contact with other infected individuals. After a certain period, they either recover, meaning they are immune and no longer spread the information, or return to the susceptible state, repeating the spreading process. Figure 2.1 illustrates the diagram of the classic epidemic models.

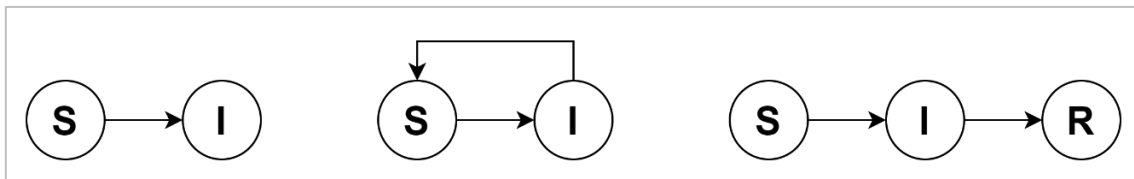


Figure 2.1: Diagram of SI, SIS, and SIR model respectively

Using the SIR model and multiparallel least-square fit for optimisation, Abdullah & Wu (2011) were the first to demonstrate that the dynamics of information dissemination on social media can be modelled using epidemic models. With the goal of maximising the spread of

information in a cost-effective manner to benefit political campaigners, Kandhway & Kuri (2014) used both the SIS and SIR models, and applied various control strategies to optimise their influence at a minimal cost. The diverse social media users speaking different languages made Wang et al. (2019) to develop a new SIR model that separates the population into distinct language groups. Through sensitivity analysis, they revealed that cross-transmission between the language groups increases the spread and influence of rumours. Li & Ma (2022) used a SIR model to investigate how propaganda and post-deletions affect information dissemination. They demonstrated that their model could optimise at a lower cost while reducing the number of spreaders. Castiello et al. (2023) proposed a very similar model to the SIR model, called the Ignorant-Spreader-Recovered (ISR) model, where they discovered that focusing on the early stages of the spread and using fewer data could accurately predict both the peak number of spreaders and the peak time occurs, compared to analysing the entire dataset.

The classic epidemic model provides a simplistic view of how information spreads by categorising the population into fixed compartments without accounting. However, it could not entirely capture the real-world dynamics, as it does not account for the complexities of individual behaviours or external factors. Han et al. (2014) noted that the dynamics of social interaction are too complex to be depicted by the simplicity of epidemic models. Because the model assumes a homogeneous behaviour among each compartment, it overlooks the differences between individuals, resulting in an inaccurate reflection of real-world user behaviour.

2.2.2 Improved Epidemic Model

From the traditional models, researchers used a variety of roles to examine how social, natural, and individual factors affect the information dissemination. They proposed better models to more accurately represent the dynamics of the spread in social media networks.

One example of an improved model would be the SEIR and SEIRS model where a person is 'exposed' (E) to the information once seen but is not sharing it yet. Han et al. (2017) created the SIRD model, which considers a 'doubt' psychological factor, and deploy a monitoring strategy using a greedy algorithm to counter the spread. Besides that, Rosli et al. (2024) incorporates an adjusting control mechanism based on the government actions aimed at reducing the rumour spread and influencing the attitude of spreaders. The model demonstrates that a sustained government effort gradually diminishes the influence of spreaders over time, effectively controlling the spread. Ojha et al. (2023) proposed the SVEIR model and find that having a 'verified' population that verifies the information before sharing would decrease the chances of spreading misinformation.

Besides the exposed state, Liu et al. (2016) proposed the SAIR model, introducing 'super-spreaders', which refer to users who have a greater ability and influence to distribute information more rapidly than an 'infected' user. By analysing the super-spreader mechanism, it can reflect the rapid and extensive spread of information that is far beyond its typical patterns, such as viral posts, compared to the conventional SIR model. Xia et al. (2023) separated the spread into its popular and levelling-off phases and use the SEIRS model combined with an ARIMA model to address limitations in the original model, which did not account for factors contributing to noise in the spread. While the hybrid model effectively captures noise trends in

the later phase, it showed discrepancies in the earlier phase, likely due to unreliable data collection methods.

Another modification to the classic model is the classification of infectors population. For example, Zhang and Chen (2019) created the SETQR model where users will spread information regardless of whether they trust or question it. With the inclusion of factors like time lag, containment and forget mechanism, the model demonstrated that these elements are crucial in influencing the dissemination process. Gopal and Kovoov (2023) proposed the SEPNS model, which categorises infectors based on whether they share rumours with positive or negative comments, analysing the impact of sentiments on the information dissemination.

To highlight, researchers have added more compartments and factors into the classic epidemic models such as verifiers, doubters, government control, and so on, to analyse the dissemination process more effectively. However, introducing additional compartments and realistic factors increases the number of parameters, which raises the complexity of the model and demands a more thorough analysis (Liu et al., 2023). Moreover, compared to the classic models, where the infected can be easily identified, it would take a lot of time and effort to identify additional classes in the dataset, as it is not possible to precisely determine how many users belong to those classes, as mentioned by Castiello et al. (2023).

2.3 Heat Energy Diffusion Model

Another approach used to study the information propagation is the heat energy diffusion model. The heat energy model describes the distribution of heat energy through diffusion within a physical system. It shows the process by which heat energy moves from hotter regions to colder regions until equilibrium is reached. The temperature gradient shows parallelism to

social media information dissemination as it resembles the varying levels of user engagements, according to Han et al. (2014).

Han et al. (2014) adopted the heat energy model to characterise the information dissemination, capturing its dynamics based on the characterisation of the rumour, its effect on individuals and the network, the accumulation of the individual's energy, and its transmission rules. The model effectively reflects the variation in infectious capabilities among different nodes and demonstrated that the spreading decreases as the appeal of the rumour and user interest diminish over time. Building on this model, Wang et al. (2019) proposed the idea of negative energy as a way to study the mitigation of rumours through injection truths, which led to an effective method to mitigate the spread of rumours in the earlier stages.

2.4 Network Model

Network modelling is utilised to represent and analyse the interactions users within the social network system. Liu et al. (2023) discovered that many researchers combined the topology of social networks with the dynamics of information dissemination to developed various networking. By representing social media users as nodes and their interactions as edges, they can have a better understanding on the mechanism and factors driving the spread of information. In recent years, researchers have proposed models that simulate the information spread by drawing parallels to natural phenomena, such as the spread of forest fires (Indu & Thampi, 2019; Kumar et al., 2021).

2.4.1 Forest Fire Model

The forest fire model is a nature-inspired approach designed to simulate the spread of fires in a forest. In this model, nodes are represented as trees that can exist in different states, either the tree is alive or can catch fire and capable of spreading the fire to other trees.

Indu & Thampi (2019) compared social networks to a forest, where users, or the nodes in the network, are like trees. Using this analogy, they developed a forest-fire model to study the interactions within the network. The model examines the likelihood of the user being influenced by the information and how far it spreads. They formulate the probability of the spread of the information based on the parameters they identified that is analogous to the spread of forest fire. After simulating the spread, their findings revealed that most of the spreading happens through retweets of others' posts. Kumar et al. (2021) expanded on this traditional forest-fire model by introducing a 'Burnt' state to represent the non-spreaders who stop sharing information after receiving it. Additionally, they formulated the spread probability as the weighted summation of each parameter. These modifications resulted in a more efficient model that outperforms existing predictive models.

2.5 Summary

In summary, this literature review has explored various models used to study information dissemination on social media, highlighting the contributions made by the researchers. Epidemic models have been widely applied to study the spread of information, with many studies focusing on controlling the spread. However, the homogeneity of these model simplifies the complexity of human behaviour resulting in inaccuracies in representing real-world social interactions as it only focuses on the spread at a population level. Therefore, alternative approaches, such as the heat energy model and network model, provide more

valuable insights into the dynamics of the information spread, offering different perspectives on individual behaviours and its interactions. While each model has its strengths and limitations, the development of understanding the information dissemination in social media continues to evolve.

Based on the insights gained from the review, the forest fire model is chosen for this project due to its ability to capture the dynamics of the spread at a microscopic level, focusing on individual users and their interactions in the social network. Additionally, given the limited number of studies exploring this approach, it could provide a contribution to the existing body of knowledge by extending the application of the forest fire model on complex networks for studying information dissemination.

In the next chapter, the methodology of the project will be detailed, outlining the process and steps used to achieve the objectives. This consists of providing an overview of the project design, probability formulation for the forest fire model, data collection methods, and validation approach used to validate the reliability of the model.

CHAPTER 3: METHODOLOGY

3.1 Introduction

This chapter focuses on the detailed flow of the methodology used in conducting the project and the step-by-step techniques where the details of how the research is carried out. Firstly, it explores the dynamic system of information sharing in social media and its resemblance to forest fires. By identifying the factors involved in the system, parameters are defined to determine the probability of the spread of the information in the network. Additionally, the chapter also outlines the data collection process required for this project, which will be used to develop a forest-fire model to study to spread of information in the network. The methodological process of this project can be visually summarised in a flowchart as shown in Figure 3.1.

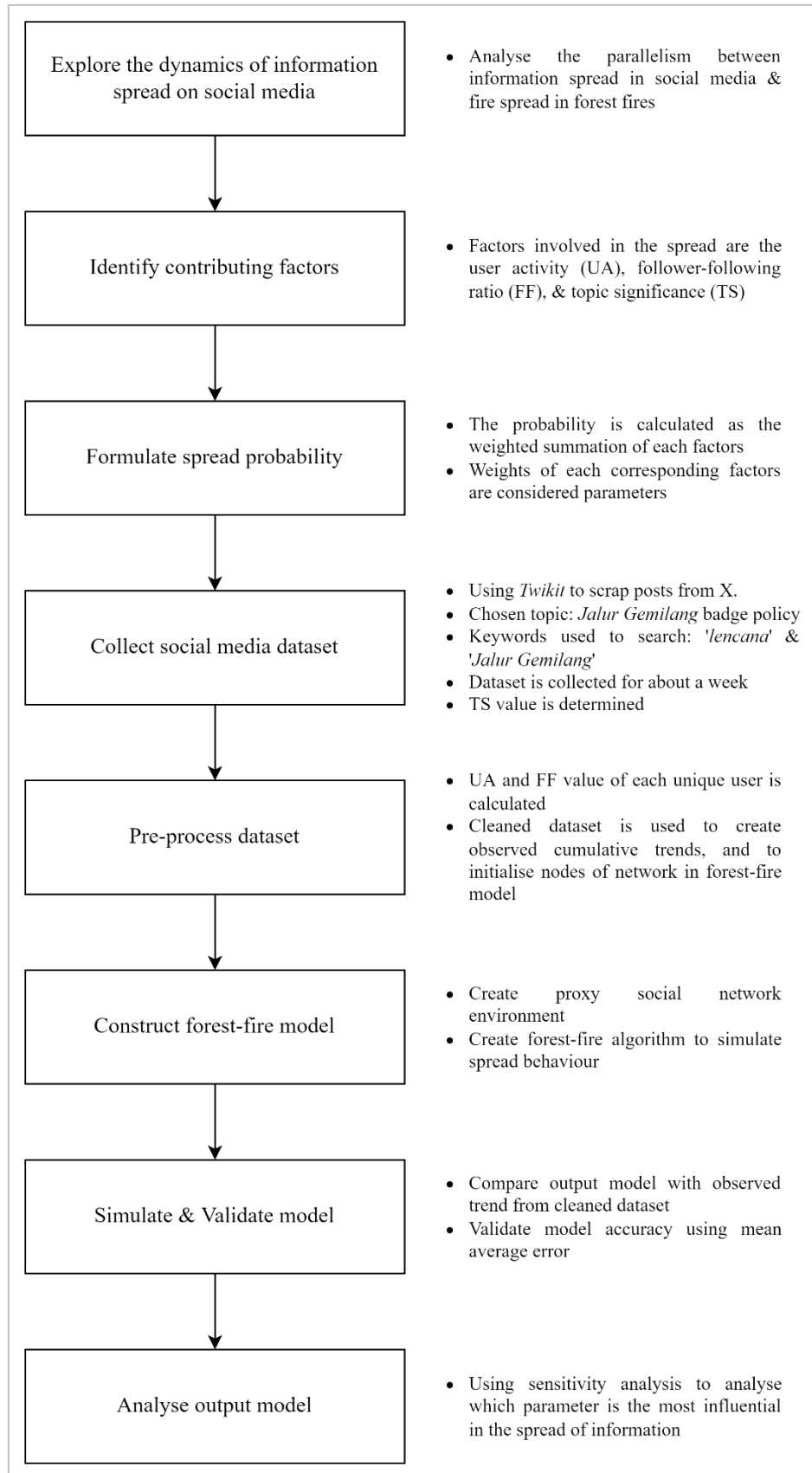


Figure 3.1: Flowchart of the methodological process of the project

3.2 Dynamics of Information Spread on Social Media

Based on Figure 3.1, the project begins by understanding how information is spread across social media networks to understand its components and interactions, as well as its parallelism to the spread of forest fires.

Information on social media spreads through a dynamic process where users in the platform share posts within their networks. Figure 3.2 illustrates the information spread in the network. Based on Figure 3.2, when someone creates a post about a topic they would like to share, it becomes visible to others, whether they are followers or public users in the network. These users can decide whether to engage with the information by reposting it or creating a new post on the same topic. This cycle of sharing increases the visibility of the information, allowing it to reach to new audiences.

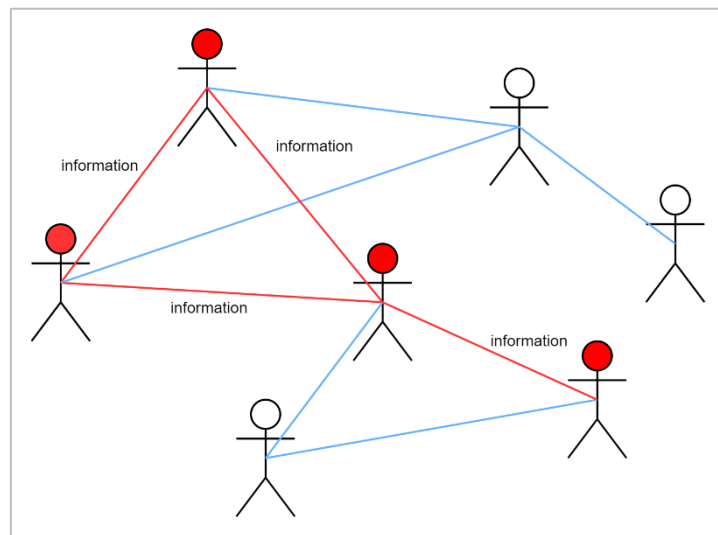


Figure 3.2: Information spread in social network

This social phenomenon resembles the spread of fire in a forest in many ways. Just as trees catch fire and spread its flames in the forest, users are influenced by the information and

share it across the social network. There are certain factors that influence the spread of the fire, and these factors can be mapped to model the spread of information in social media. For example, the density of the forest, the wind velocity, and the topography of the forest area are key factors that determine the probability of the tree catching fire according to Indu & Thampi (2019). Similarly, factors such as the follower-following ratio of user, the topic significance of the information, and the user's account activity can affect the sharing of information on social media. The parallels between the factors in forest fire and information sharing are shown in Table 3.1.

Table 3.1: Parallels between forest fire spread and information spread

Factors affecting forest fire spread	Factors affecting information spread
Density of forest	Follower-following ratio of user
Wind velocity	Topic significance of information
Topography of forest area	User account activity

Based on Table 3.1, in a forest, fire spreads faster when the trees are densely packed as they can ignite one another rapidly. Similarly, on social media, users with a larger follower-following ratio can reach out information to a broader audience, increasing the chances of rapid engagement. Besides that, wind makes fires spread faster by intensifying the flames. In social media, a trending or highly engaging topic acts like the wind, propelling information into broader discussions. This captures the users' attention from all over the world which resulted them to post their statuses on that topic (Kumar et al., 2021). Lastly, the topography of a forested area refers to the shape of the land, which influences how the fire spreads affecting its direction and intensity. Indu & Thampi (2019) relates the topography to the user's social media

activity. Active users with a lot of engagement can significantly amplify the reach and visibility of their posts.

3.3 Modelling

In this section, the probability of the user spreading information (p) is determined based on the contributing factors and components identified within the system. Once the probability is determined, it will be used in a proposed forest-fire algorithm to analyse the spread of information in the social network.

3.3.1 Defining the Spread Probability, p

After analysing the dynamics of the spread of information, it can be assumed that the spread probability, p , is influenced by the factors, which are the user's social media activity, the significance of topics shared, and the user's follower-following ratio.

3.3.1.1 User Activity (UA)

Adapted from Kumar et. al (2021), the user's social media activity can be determined by considering the total number of posts they, the spreader, have posted within a specific period of time, as shown in Equation 3.1. The period of time will be the duration for which the X data is collected, which is measured in days. A higher UA value indicates that the spreading user is very active posting a lot of content on X, thus increasing the visibility and likelihood of other users to engage with the contents through sharing or creating new posts.

$$UA = \frac{\text{Total number of tweets posted by user during data collection period}}{\text{Duration of the data collection period (days)}} \quad \text{Equation 3.1}$$

3.3.1.3 Follower-following Ratio (FF)

The follower-following ratio of a user determines the influence of a user in the social network platform. Based on Indu & Thampi (2019), FF is calculated by dividing the number of accounts following the user with the number of accounts followed by the user as shown in Equation 3.2. According to them, the higher the FF value, the more influential the account holder is, whether they are a celebrity, an organisation, or a renowned institution. Because of how highly regarded they are by the public, albeit respectable or controversial, it creates a cognitive bias towards authority, in other words an authority bias, guiding them to engage with the information regardless the type of message they deliver. In a research study conducted by Milhazes-Cunha & Oliveira (2023), they discovered that authority figures are responsible for creating online discussions, where supporters will actively engage in the discussion sharing their thoughts and beliefs.

$$FF = \frac{\text{Number of accounts following the user}}{\text{Number of accounts followed by the user}} \quad \text{Equation 3.2}$$

3.3.1.2 Topic Significance (TS)

In general, users typically tweet about topics that are popular, aiming to attract attention from other users in the network. Proposed by Kumar et al. (2021), a ranking score of the topic

significance is illustrated in Table 3.2. According to them, the higher the score is, the higher the likelihood the user tweeting about the topic.

Table 3.2: Ranking of scores based on topic significance

Topic Significance (Rank)	Global interest	Local interest
Trending	4	3
Non-trending	2	1

According to Table 3.2, Kumar et al. (2021) stated that topics of global interest will be shared from many users across the globe, resulting in a broader reach compared to topics of local interest, which focus on users within a specific region, such as a country. Besides that, trending topics, which are currently receiving significant attention, tend to attract a lot of engagement because of its high visibility on social network platforms compared to non-trending topics, which often attracts a niche group of users and are less likely to go viral.

To assign the TS values, the rankings are quantified into corresponding ranges between 0 to 1, which are used to value the importance of a topic based on its rankings. The TS value will be randomly selected for the calculation of p from the respective range as shown in Table 3.3 to avoid bias and reflect the stochasticity of the spread.

Table 3.3: Range of TS values according to ranking

Ranking	TS ranges
1	[0.01-0.25]
2	[0.26-0.50]
3	[0.51-0.75]
4	[0.76-1.00]

After determining the UA, TS, and FF values, the spreading probability, p , for each non-spreader after receiving the information, is calculated as the weighted summation of those three factors, which is shown below.

$$p = (w_1 \times UA) + (w_2 \times FF) + (w_3 \times TS) \quad \text{Equation 3.3}$$

$$w_1 + w_2 + w_3 = 1 \quad \text{Equation 3.4}$$

The utilisation of weightages to determine the probability is supported by Kumar et al. (2021) and Mittal et al. (2019) who also applies weightages in formulating their outcome of the interaction of their models. Based on the Equation 3.3, w_1 , w_2 and w_3 represent the constant weights associated with the respected component factors, which are user activity (UA), topic significance (TS), and follower-following ratio (FF). The sum of all the weights is equal to 1, as shown in Equation 3.4. Because the weights control the magnitude of each factor which determines the success of a spread, they are considered to be the parameters responsible for the model's behaviour. Therefore, the parameters are defined as how non-spreaders perceive the

importance of each factor when deciding whether to spread the information. Table 3.4 depicts the definition of the parameters.

Table 3.4. Parameter definition and its ranges

Parameter	Definition	Ranges
w_1	The weight reflecting the influence of UA on the spread	[0, 1]
w_2	The weight reflecting the influence of FF on the spread	[0, 1]
w_3	The weight reflecting the influence of TS on the spread	[0, 1]

3.3.2 Data Collection

The data collection processes which include X's scraping and data preprocessing, to provide relevant data to inform and validate models necessary for accurate predictions.

3.3.2.1 Dataset Description

The proposed spreading model will be tested on real social media network from X. To ensure the relevance of the simulation, posts are collected around a topic that was actively discussed at the time. The topic chosen for the project is regarding the Malaysian Ministry of Education's announcement to implement *Jalur Gemilang* badges for all school students nationwide (Bernama, 2025). Following the announcement, it sparked a lot of discussions online, with reactions ranging from strong support to criticism and neutral commentary, as shown in Figure 3.3. As a result, this makes a relevant case study to model real-world information dissemination.



Figure 3.3. Example posts from X regarding the *Jalur Gemilang* badge policy

Using Python's *twikit* library, the posts related to the topic is collected using the keywords, '*lencana*' and '*Jalur Gemilang*', as these terms are frequently mentioned in social media following the policy announcement. Because of this, the dataset collected are limited to posts in Malay and English to ensure semantic relevance and to minimise possible noise in other languages. The posts are scraped about a week, from 27th March 2025 to 1st April 2025, which was when the policy was initially announced to the point when it is no longer gain any public interest. The collected dataset included the following attributes for each post:

- i. `username`: The handle of the user who published the post.
- ii. `follower_count`: The number of accounts that follow the user.
- iii. `following_count`: The number of accounts the user is following.
- iv. `text`: The textual content of the post.
- v. `date_created`: The timestamp of when the post was created.

After scraping, the dataset contains a total of 1373 posts and will be cleaned and prepared before using it in the model and for validation.

3.3.2.1 Data Scrapping

The cleaning and preparation of the dataset is carried out using Python's *pandas* library. The cleaning process involved the following steps which is illustrated as a flowchart as shown in Figure 3.4.

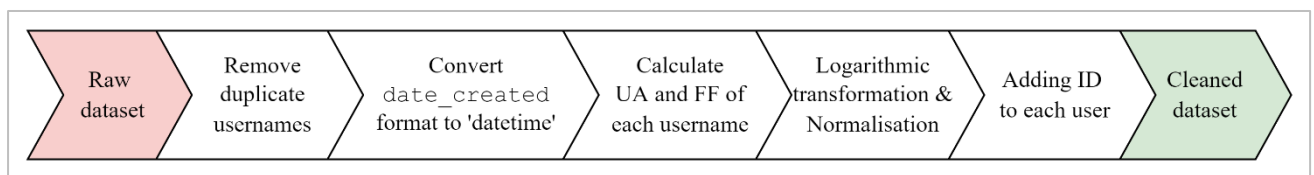


Figure 3.4. Flowchart of cleaning the raw dataset

Firstly, a `count` attribute is added to count the frequency of unique usernames in the dataset. Then, posts with identical usernames are removed to avoid data redundancy. After removing the duplicates, there is a total of 949 unique usernames in the dataset. After that, the datatype of `date_created` is converted from a string to a datetime format because it will be used to categorise the time for creating the observational cumulative trend of the dataset. Since the data was collected over a week, the observation period is relatively short to investigate the daily trend. Hence, each observed day is segmented into thirds (eight-hour intervals), for instance, which provides sufficient detail to observe the spread within the short timeframe. From this, each post is categorised by its day and corresponding time interval, ranging from the first interval of the first day to the last interval of the final day (e.g., Day 1-Third 1, Day 7-Third 3).

The next step is to obtain the UA and FF values of each user. To calculate the UA values, simply dividing their `count` values with seven. As for the FF values, the values from `follower_count` are divided with the values from `following_count`. However, there are users who either have 0 followers or 0 followings, or both. These would lead to a division-by-zero errors and result in FF value of zero to some user implying that they have no influence in the spreading of information. This is not always accurate as accounts with zero followers can still influence others through reposts, albeit with limited visibility. To address this, all follower and following values of 0 were replaced with 1. Replacing the values has little effect on the overall spread and does not significantly compromise the integrity of the dataset as the change is very small and uniform across all affected entries.

To ensure the UA and FF values are on a comparable scale and neither disproportionately outweighs one another, normalisation is applied to both. However, before normalisation, logarithmic transformation is performed first to reduce the extreme gap between small and large values. This reduces the skewness hence, making the data more suitable for scaling. Finally, the values are normalised to the range (0, 1). The reason for not normalising to [0, 1] is to avoid having extreme interpretations to the boundary values, as a value of 0 implies no influence, while a value of 1 implies complete influence, both of which are unrealistic in the context of the model.

Lastly, user IDs are added to each user starting from 0. Using the cleaned dataset, a cumulative frequency list of the total number of spreaders over time is created to observe the information spread over time and to validate the forest-fire model. Additionally, the dataset is also be used to initialise the nodes for the creation of the social network environment, and to simulate their interactions via the forest-fire algorithm.

From the cleaned dataset, it will be used to do the following:

- i. Creating observational cumulative trends of the number of spreaders over time for model validation purposes.
- ii. Initialising nodes and edges for the creation of the social network environment, which is used to simulate the information spread via the forest-fire algorithm.

3.3.3 Forest-Fire Model

To simulate the spread of information among users in social media, the project uses an agent-based modelling approach using the forest-fire model. The forest-fire model, originally developed to simulate the spread of fires in forests, is adapted in this project to represent the probabilistic spread of information between users in a social network. The development of the proposed forest-fire model is comprised of two parts, which is creating a proxy social network environment and creating the forest-fire algorithm.

Using the cleaned dataset, a social network environment is created where each node represents the existing user in the dataset, similar to ‘trees’ in a forest, and each edge represents the potential paths for information transmission. The forest-fire algorithm then simulates how information spreads through the network. Spreaders of the social network are considered as ‘fires’, for they are posting information online. At each step of the simulation, the spreaders have a probability of ‘igniting’ their neighbours, which is influenced by the contributing factors, as defined in Equation 3.3.

3.4 Summary

This chapter summarises the methodological process for this project. It begins with studying the information spread dynamics in social media and how it is similar to forest fires. Next, factors such as user activity, follower-following ratio, and topic significance, are identified to formulate the probability of a non-spreader spreading the information according to the factors. The spread probability is defined as the weighted summation of the factors using constant weights corresponding to the factors as parameters. Social media dataset from X website is scraped and pre-processed so that the cleaned dataset will be used for the creation of the observational trends of the number of spreaders over time, and the creation of the forest-fire model.

In the next chapter, the proposed forest-fire model is constructed to achieve the project's objectives. Additionally, sensitivity analysis will be performed on the model to investigate the influence of each parameter to the spread dynamics of the model.

CHAPTER 4: IMPLEMENTATION & ANALYSIS

4.1 Introduction

This chapter will focus on explaining the construction the social network environment, the forest-fire algorithm, and the sensitivity analysis of this project. The execution of the simulations is performed on a system with configuration: 16 GB RAM, 11th Gen Intel(R) Core(TM) i5-11400H CPU @ 2.70GHz (6 CPUs), 2.69GHz processor. These sections will be carried out in order to proceed with the results.

4.2 Environment Setup

Before conducting the simulation, an agent-based environment must first be established. Therefore, the social network environment is constructed using Python's *networkx* library.

In the network, the nodes represent the users in the cleaned dataset. For each user in the dataset, the nodes are created with attributes such as the user ID, the normalised UA and FF values, as well as the *Fire* status of the user, which is initially set to false. After that, the initial spreaders in the network are assigned based on the cumulative frequency list, with their *Fire* status of the spreaders set to true.

Due to privacy constraints, the users' social connections, which are their follower and following links, are not publicly accessible. Therefore, an alternative approach is made to connect the users. In this model, the connections between the users are established based on co-occurrence, where users who post within the same time window are assumed to be linked. For example, if two users post a topic within an hour for instance, they are considered

connected to each other. In other words, it defines how close in time the users must post in order to be considered connected in the network.

The decision to connect them based on a time window reflects on common user behaviour in social media, where users are more likely to view and respond to posts that are recent to stay updated with the current news and trends. Because of this, users who engage with the same topic within the same timeframe are considered to be in a shared information environment, where they may influence one another due to the temporal visibility and relevance of the content. This proxy network will be used to simulate the information spread using the forest-fire algorithm.

4.3 Forest-fire Algorithm

The proposed forest-fire algorithm intends to find out the average number of spreaders in the social network over time. To address the stochasticity of the spread within the model, a Monte Carlo simulation approach is implemented by running the simulation multiple times. For each cumulative frequency per run, the average of each element will be calculated to reduce the effects of randomness and provide a more stable estimate of the model's behaviour.

Input: Social network, $G = (V, E)$, parameter values = (w_1, w_2, w_3)

Output: Average number of spreaders over time, L_{ave}

```
1  Initialise list of cumulative frequencies,  $L = \{\}$ 
2  Repeat 50 times:
3      Initialise list of Fire nodes,  $F = \{\}$ 
4      Initialise cumulative frequency of Fire,  $C = \{\text{length}(F)\}$ 
5      Repeat  $n$  times (where  $n$  = number of segments of the dataset):
6          Initialise list of new Fire nodes,  $F_{new} = \{\}$ 
7          for each node in  $F$ :
8               $UA = \text{node}['ua']$ 
9               $FF = \text{node}['ff']$ 
10             if neighbour is Tree:
11                  $p = (w_1 \times UA) + (w_2 \times FF) + (w_3 \times TS)$ 
12                 if  $p \geq \text{threshold}$ :
13                     neighbour['status'] = 'Fire'
14                      $F_{new} = F_{new} \cup \{\text{neighbour}\}$ 
15                 end if
16             end if
17              $F = F \cup F_{new}$ 
18              $C = C \cup \text{length}(F)$ 
19         end for
20     end repeat
21      $L = L \cup C$ 
22 end repeat
23 Average each element in  $L$ 
```

Figure 4.1. Forest-fire algorithm

The forest-fire algorithm is described in Figure 4.1. Here, the algorithm takes in inputs, which are the values of the parameters, which are the weights of the features (w_1 , w_2 , and w_3), and the set of nodes in the social network environment, $G = (V, E)$, where V is the set of nodes or users in the cleaned dataset and E is the set of edges or connections that the users interact

between them. For every user of the network has these characteristics: the user ID, their UA and FF values, and their status whether they are Fire or Tree. Users that have created a post on social media are considered as Fire, otherwise they are considered as Trees. Whenever a node is on Fire, they will spread its fire to the Trees, implying that they have successfully spread its influence to spread its influence on users to create a post. The fire will keep on spreading until it covers the entire network forest.

The algorithm begins by creating a list (L) to record the cumulative frequencies of spreaders of each run, allowing them to be averaged to produce the final result which is the average cumulative frequency (L_{ave}). Next, for every simulation run, two lists are initialised, which is a list to record the ID of users that have the Fire status (F), and a cumulative frequency list of each Monte Carlo run (C) with the first value already assigned as the length of F , implying the number of spreaders of the first timestamp.

The spreading phase begins through iterations corresponding to the number of time intervals segmented in the cleaned dataset, denoted as n . A new list (F_{new}) is created to record the ID of users that are newly infected. Each node in F is then iterated and its UA and FF values are obtained. Then, the algorithm will visit every neighbour of the node. If the status of the neighbour is Tree, then the spread probability, p is calculated for the neighbour using Equation 3.3. The TS value is chosen at random from the selected TS range from Table 3.3. Based on the collected dataset, the topic is considered to be trending and local (rank 3), hence the TS value will be randomised value between range $[0.51, 0.75]$ at every iteration. If p surpasses a threshold value, then the neighbour is considered to have created a post and thus, has changed its status to Fire and the neighbour is then added to F_{new} . Here, the value of the threshold is considered random to simulate the random behaviour of users. After visiting every neighbour,

the nodes from F_{new} are added into F and the cumulative frequency, C is updated. This process is repeated for n times, capturing the dynamics of the spread from the first to the last timestamp.

After obtaining the final number of C , the spreading phase has ended. C is added into L and the simulation repeats until it has run for a total of 100 simulations. Finally, when L has obtained 50 cumulative frequency lists, the algorithm averages each element in the list and outputs a list of average number of spreaders over time (L_{ave}).

4.4 Sensitivity Analysis

To identify the influence of each contributing factor on the spread of the model, a global sensitivity analysis was conducted using Sobol sensitivity analysis (or Sobol's method). This method quantifies how each parameter contributes to the variance of the model output by sampling the model from a wide combination of parameter values. Sobol's method was implemented using Python's *SALib* library.

The Sobol method was conducted on the three weighted parameters (w_1 , w_2 , and w_3) that control the influence of the contributing factors, as they determine the probability of the spread between users in the model. Saltelli's sampling method was used to generate a range of weight combinations within the bounds $[0, 1]$ for each parameter. The number of samples has to be sufficiently large to for a more reliable and accurate sensitivity analysis. Next, the weights were then normalised so that the sum of the weights is always one. For each combination of weights, the model was executed, and the final number of informed users was recorded as the output metric for sensitivity evaluation.

The result output from the analysis produced the first-order (S_1) and total-order (S_T) Sobol indices. The first-order index measures the contribution of a single parameter to the

variance output, isolating its individual effect, and the total-order index, in contrast, measures the parameter's overall influence, including its interactions with other parameters. These Sobol results are used to determine which of the following factors has the greatest influence on the spread of information.

Due to the device's limitations to execute a large number of simulation runs, the process was parallelised by CPU using Python's *mpi4py* library. Using an embarrassingly parallel structure, the parameter samples were distributed across multiple processing cores to significantly reduce the computation time.

4.5 Summary

This chapter explains the construction of the information spreading model and the sensitivity analysis applied to evaluate it. Firstly, a proxy social network environment is created as a foundation for the simulation. Next, the simulation is executed using the forest-fire algorithm, which models the behaviour of each user in the network through the spread probability. Lastly, Sobol's sensitivity analysis is implemented to assess the influence of each parameter on the output variance of the model, which will then be used to identify the most impactful factor in the spread of information. The simulation and sensitivity analysis results will be presented in the next chapter.

CHAPTER 5: RESULTS & DISCUSSION

5.1 Introduction

This chapter discusses the results of the forest fire model. Additionally, it also discusses the effect of parameters on the output variation of the model based on Sobol sensitivity analysis.

5.2 Model Results

Figures 5.1-5.3 shows the cumulative distributions of the number of spreaders over time in the cleaned dataset to which the observed days are segmented into eight-, six-, and twelve-hour intervals (or segments) respectively.

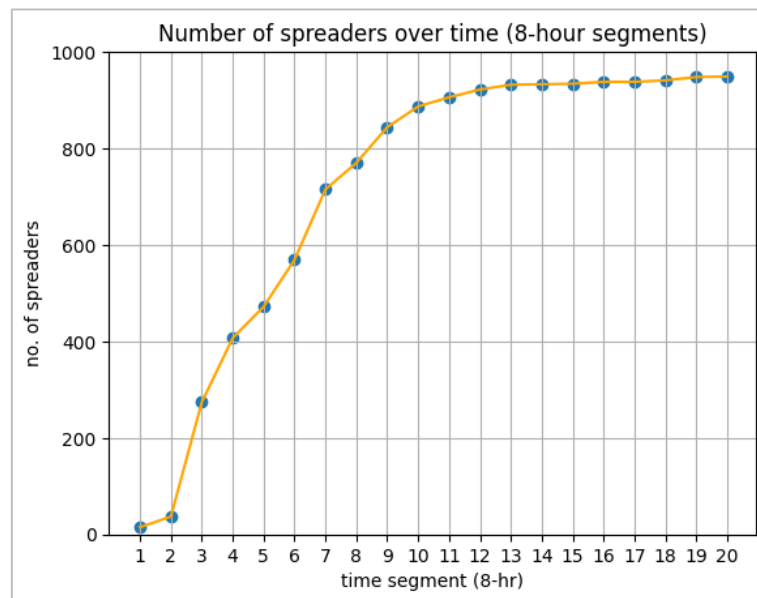


Figure 5.1. Trend of number of spreaders over eight-hour intervals

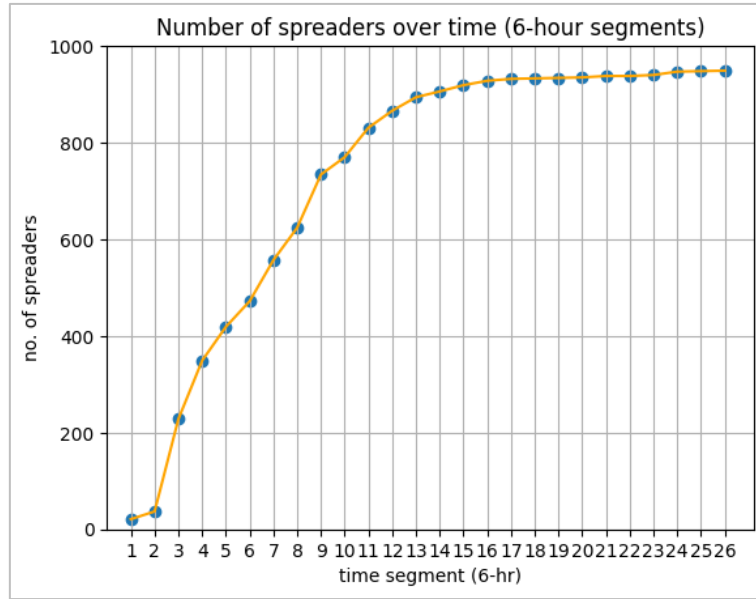


Figure 5.2. Trend of number of spreaders over six-hour intervals

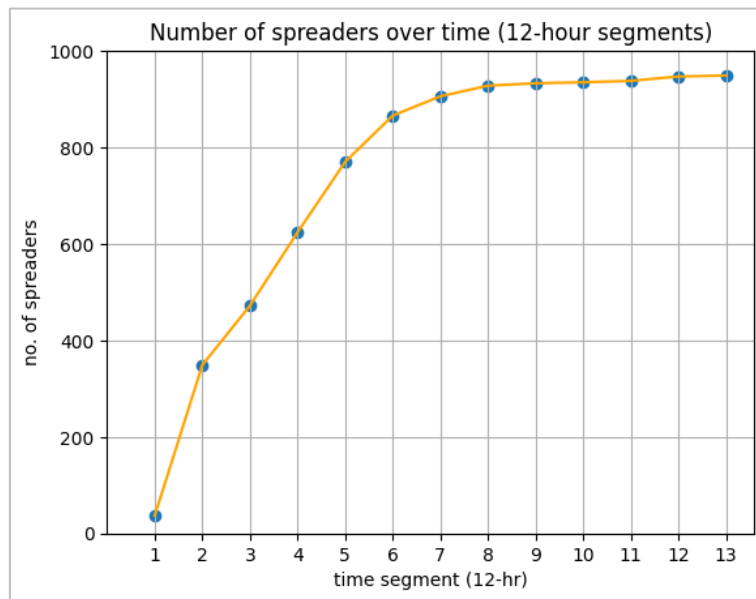


Figure 5.3. Trend of number of spreaders over twelve-hour intervals

To determine the flexibility of the model, several tests were conducted using different network environments. Specifically, the simulation was executed with multiple networks of varying time window connections, which are eight, six, and twelve hours. The values of the

weighted parameters are equally distributed, where $w_1 = \frac{1}{3}$, $w_2 = \frac{1}{3}$, and $w_3 = \frac{1}{3}$. Figures 5.4-5.6 illustrate the comparison between the models' information spread over time under different observed time segments. Table 5.1 tabulates the mean average error (MAE) of these results.

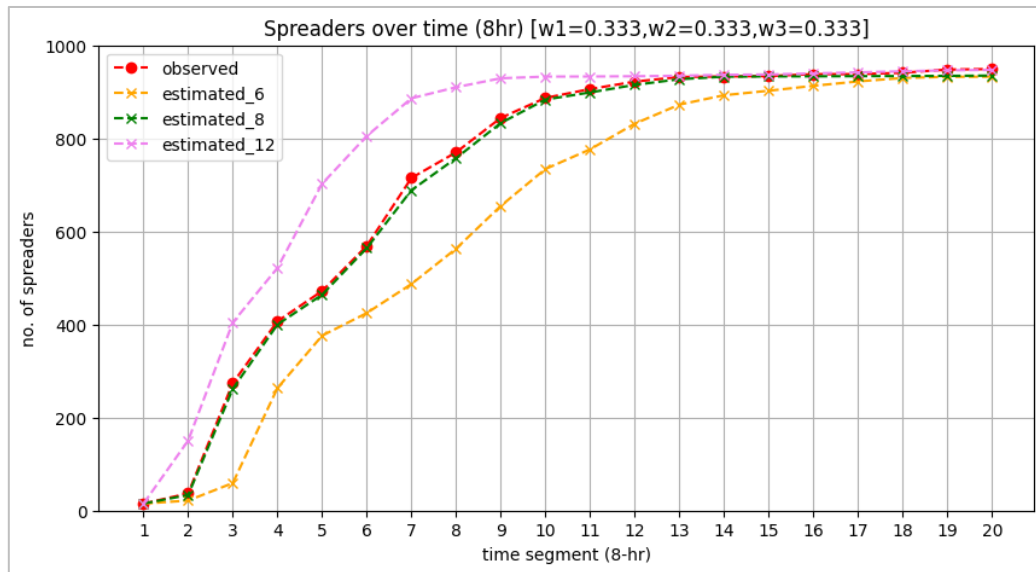


Figure 5.4. Model results observed under 8-hour segments

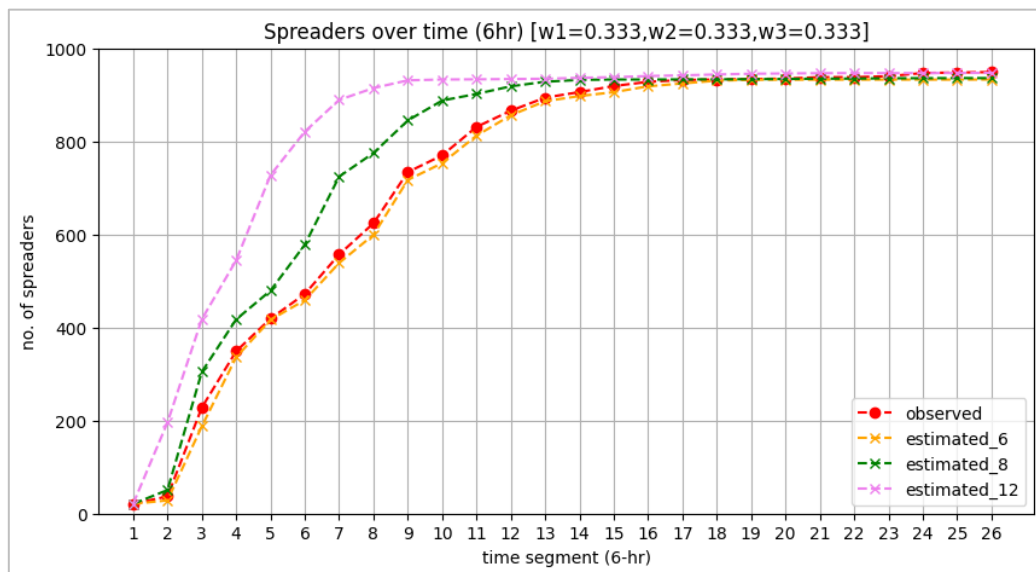


Figure 5.5. Model results observed under 6-hour segments

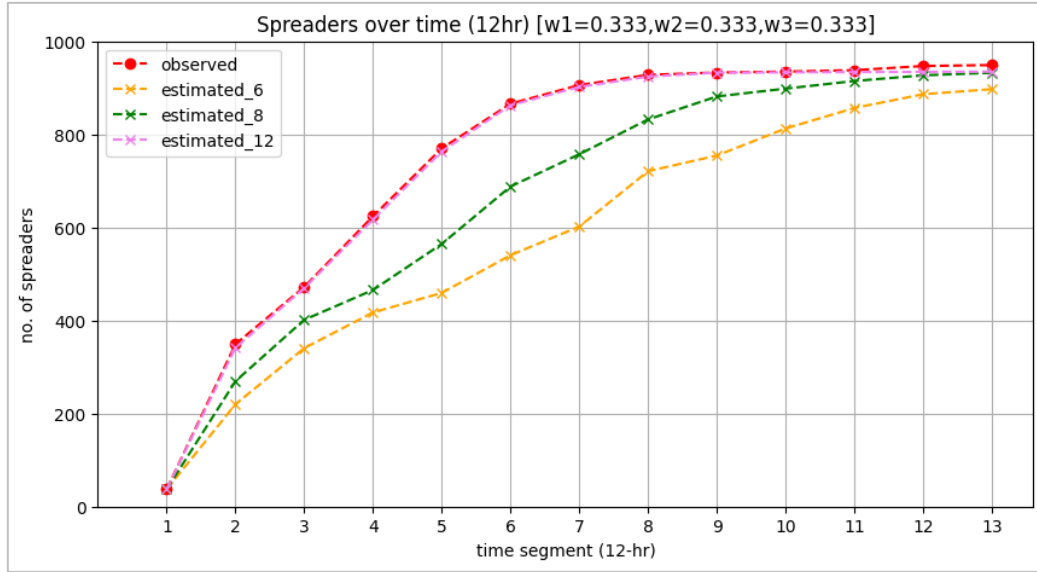


Figure 5.6. Model results observed under 12-hour segments

Table 5.1. MAE of each model result

		Time segment			
Time window		6	8	12	
	6	11.4846	91.31	162.2154	
	8	43.2385	7.62	83.4	
	12	97.2154	66.39	5.2154	

Based on the results, it is observed that the model gives more accurate outputs when the network's time window connection is synchronous, or is aligned, with the time segment of the observed dataset. When the time window covers the same duration as the dataset's time segment, it implies that the user interactions within the network occur within the same timeframe. In contrast, a misalignment between the two can result in unrealistic spread paths because users may still be active spreading information even after the simulation has move to

the next iteration, or users have become inactive while model is still observing. Therefore, when both time settings are matched, the network structure mirrors the users' behaviour in the dataset resulting in a more faithful simulation of the information spread.

Using a time segment of eight hours as the observational trend, Figure 5.7 shows an arbitrary output of the model where the combination of weights ($w_1 = 0$, $w_2 = 0$, and $w_3 = 1$) have the lowest MAE value (5.37). Based on the combinations, it shows that the spreading of the *Jalur Gemilang* badge policy is primarily driven by the topic significance, suggesting that the users are most influenced by the relevance of the implementation of the policy (TS) rather than the spreader's online activity levels (UA) or their social connectivity (FF).

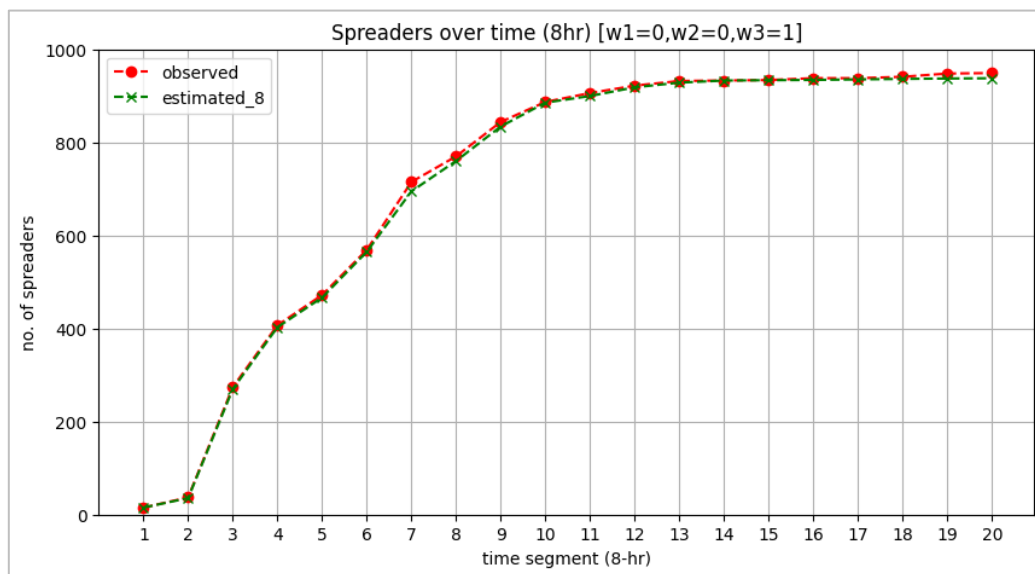


Figure 5.7. Output model with the most optimal weight combination

5.3 Sensitivity Analysis Results

Similarly, Sobol sensitivity analysis was also carried out on each variation in the time configuration, resulting in a total of 9 Sobol results. The reason of performing multiple analyses is to investigate whether the change in the time segments and time window will influence the sensitivity of the model and the contribution of each parameter. The first-order (S_1) and total-order (S_T) Sobol indices with confidence intervals, under different configurations are tabulated in Table 5.2.

Based on Table 5.2, it is observed that the results of the analyses are consistent, whereby w_3 has the largest effect, followed by w_1 , and w_2 with the least value. The consistent ranking of the parameters indicates that the relative influence of these parameters remains stable across different time segmentation and time window configurations albeit varying magnitudes. Besides that, some results have negative Sobol index values or large confidence intervals. While these do not mean a negative effect on the model, these are likely due to insufficient number of samplings or model evaluations. With only 50 Monte Carlo runs per Sobol sample due to hardware limitations, it might not be enough to fully average the stochastic noise in the simulation, which may affect the accuracy of the Sobol results.

Focusing only on the relative influence of the weight parameters, w_3 , which corresponds to the weighted input of the topic significance factor, consistently shows the highest S_1 and S_T values. This suggests that, within the model, the spread of the information is primarily driven by the degree to which topic significance is emphasised through its assigned weight.

In other words, users are likely to spread information based on how much importance they place on the topic, regardless of how active the netizens posts (from user activity) or how

many connections they have (from follower-following ratio). The topic of a content is what motivates the users to share. This aligns with many real-world observations, where content relevance outweighs social metrics in driving engagement. For example, in a literature review conducted by Schreiner & Riedl (2019) in determining the factors of social media engagement, they discovered that a majority of the searched papers focused on the context of the content compared to other factors.

5.4 Summary

This chapter explains the results of the model and the sensitivity analysis. Based on the model results, it is shown that the accuracy of the model depends on the synchronisation between the time segmentation of the dataset and the time window of the network's connectivity. From the Sobol sensitivity analyses, it is observed that the relative influence of each weight parameters is consistent across different network configurations albeit varying magnitudes. Additionally, it can be deduced that the spread of the model is heavily influenced by the degree corresponding to topic significance. This means that information is spread based on how users perceive the content, regardless of popularity in social media.

Table 5.2. Sobol results under different network configurations

TW	6			8			12		
	w_1	w_2	w_3	w_1	w_2	w_3	w_1	w_2	w_3
Time segment = 8 hours									
S_1	0.05922625	0.03496174	0.8863871	0.16113986	0.13446603	0.53190127	0.13346978	0.09684021	0.78939451
Slconf	0.07728847	0.05527915	0.32835136	0.13092423	0.11980148	0.18643208	0.09370919	0.08641872	0.24935743
S_T	0.0775896	0.04727031	0.92117446	0.25090864	0.21522828	0.62080925	0.15719155	0.12823	0.8550864
STconf	0.03410266	0.0252661	0.18098987	0.07507705	0.05672461	0.15429479	0.04917049	0.04158884	0.20048189
Time segment = 6 hours									
S_1	0.21359933	0.14536291	0.67655706	0.0679407	0.03422487	0.83660821	0.18167776	0.13356516	0.59130159
Slconf	0.10324615	0.1049037	0.19296468	0.08064208	0.05667386	0.26168004	0.13838407	0.10234841	0.17721161
S_T	0.2619088	0.20621651	0.64229105	0.09848592	0.06901973	0.85723289	0.27175238	0.20298178	0.63178172
STconf	0.12389677	0.08686036	0.15839744	0.03912842	0.0321889	0.19039864	0.09771498	0.08353057	0.1952544
Time segment = 12 hours									
S_1	-0.0221166	-0.0250224	0.8509538	0.09824724	0.05979408	0.82598233	0.07810672	0.02238108	0.90359203
Slconf	0.09644768	0.09711524	0.42492888	0.08434961	0.07500114	0.24229815	0.08084719	0.06160225	0.28379406
S_T	0.22405419	0.18537219	1.04491641	0.13413976	0.0987981	0.85845844	0.12755473	0.10740795	0.87197213
STconf	0.1153317	0.10388211	0.28814161	0.04412247	0.05233255	0.2171079	0.059547	0.04568979	0.20693211

(Keys: TW = time window; Slconf = S_1 confidence interval; STconf = S_T confidence interval)

CHAPTER 6: CONCLUSION & FUTURE WORKS

6.1 Introduction

This chapter examines the results of the project by comparing the objectives with its corresponding achievements. Additionally, the limitations of the project are also discussed as well as list of improvements for future works.

6.2 Achievement of Objectives

Table 6.1 summarises the objectives as stated in Chapter 1 and its corresponding achievements. Based on the table below, the information spread is successfully modelled after the proposed forest-fire algorithm. The algorithm is based on a spreading probability, which is weighted summation of the contributing factors: user activity, follower-following ratio, and topic significance. Given that the weights are used as parameters of the model, using Sobol sensitivity analysis, it is observed that the spread is mostly influenced by the degree corresponding to topic significance. Based on these results, the objectives are fulfilled.

Table 6.1. Objectives and its corresponding achievements

Objective	Achievement
To create a computational model that reflects the dynamics of information spreading on social media.	A proxy social network environment is created as a foundation for the simulation based on the proposed forest-fire algorithm.

To validate the predicted model against real X data to assess its predictive accuracy.	The accuracy of the model depends on the configuration of the network, which means the time segmentation of the dataset has to be synchronous with the time window of the network's connectivity.
To investigate the impact of each identified parameter on the information spread through sensitivity analysis.	Using Sobol sensitivity analysis, the spread of the model is dominantly influenced by the weight corresponding to the topic significance of the post.

6.3 Project Limitations

Several problems were encountered throughout the completion of this project. Firstly, one major problem lies of the construction of the social network environment. Limited research was done in finding how the users of the network are linked to each other, as many articles explained how their network is created but not connected. The manner in which the users are connected is based on simplified assumptions drawn from personal experience of social media usage rather than grounded in existing literatures, thus may not accurately reflect the complex and dynamic structure found in real-social networks. As a result, the model lacks flexibility when applied to broader scenarios, as its predictive strength only works in specific conditions.

Another limitation is the hardware resources used to perform the simulation and sensitivity analysis. The accuracy of the Sobol indices can be improved by increasing the number of Monte Carlo iterations and Saltelli samples. However, due to hardware limitations and the absence of applying GPU-based parallelisation, computations were executed using

CPU parallelisation through the *mpi4py* library. Even though this approach significantly reduced the computation time, it is still not as efficient as GPU parallelisation, which could potentially allow for higher sample counts and more robust analysis.

Lastly, time constraints also posed a challenge throughout the project. While fundamental concepts behind Sobol sensitivity analysis were understood and implemented, limited time prevented a deeper understanding of advanced aspects of the technique. With more time, a more comprehensive discussion of the sensitivity results could have been developed, potentially uncovering subtler interactions and dependencies within the model parameters.

6.4 Future Works

For future research and development of this project, there are a number of improvements that can be done to improve the model's capabilities. Several aspects can be considered, such as employing GPU parallelisation to run more samples for better Sobol result.

Besides that, given that topic significance played an important role in the spread of the model, sentiment analysis of the collected dataset can be applied to assess whether the spread is influenced by emotions.

Next, a more data-driven approach to construct the network can also be incorporated to better capture how users are actually connected, such as using comments, mentions, and retweeted relationships from X.

Lastly, the model can be enhanced by incorporating more diverse user behaviours and interactions that better reflects the complexities of the real social network, such as popularity decay, or even user's resistance to spread.

6.5 Summary

This project aimed to model the spread of information on social media using a forest-fire model approach and to assess the influence of different contributing factors through sensitivity analysis. Findings from sensitivity analysis suggests that users' perception in the topic significance of the content consistently plays the most dominant role in determining whether information is shared. While the model provides useful insights, its flexibility is limited by assumptions in the network construction and computational constraints. With more resources and time, the model could be enhanced by incorporating more comprehensive approaches and behaviours that better reflects the complexity of the real social network. Overall, the project successfully demonstrates how information is spread online, thus providing a groundwork for more detailed future studies.

REFERENCES

- Abdullah, S., & Wu, X. (2011). An epidemic model for news spreading on Twitter. *2011 23rd IEEE International Conference on Tools with Artificial Intelligence*, 163-169.
- Al-Quran, M. W. (2022). Traditional media versus social media: Challenges and opportunities. *Technium Romanian Journal of Applied Sciences and Technology*, 4(10), 145-160.
- Bernama (2025, March 27). *MOE to implement Jalur Gemilang badge on student uniforms on 21 April*. Awani International. <https://international.astroawani.com/malaysia-news/moe-implement-jalur-gemilang-badge-student-uniforms-april-21-514171>
- Castiello, M., Conte, D., & Iscaro, S. (2023). *Using epidemiological models to predict the spread of information on Twitter*. *Algorithms* 2023, 16(8), 391.
- Daley, D. J., & Kendall, D. G. (1964). Epidemics and Rumours. *Nature*, 204, 1118.
- de Arruda, G. F., Rodrigues, F. A., & Rodríguez, P. M., Cozzo, E., & Moreno, Y. (2018). A general Markov chain approach for disease and rumour spreading in complex networks. *Journal of Complex Networks*, 6, 215-242.
- Dey, P., Pyne, S., & Roy, S. (2017). Information spreading in online social networks: A case study on Twitter network. *Proceedings of the 18th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, 34, 1-2.
- Gopal, G. N., & Kovoov, B. C. (2022). Modelling and stability analysis of a rumour propagation model with sentiments as microscopic observation. *Journal of Information & Knowledge Management*, 22(2).

- Han, S., Shi, Z., Zhuang, F., & He, Q. (2014). Energy model for rumor propagation on social networks. *Physica A: Statistical Mechanics and its Applications*, 394, 99-109.
- Han, Q., Miao, F., & Fan, W. (2017). Rumor spreading and monitoring deployment in online social networks. *2017 IEEE 17th International Conference on Communication Technology (ICCT), 2017*, 1347-1351.
- Indu, V., & Thampi, S. M. (2019). A nature-inspired approach based on forest fire model for modeling rumor propagation in social networks. *Journal of Network and Computer Applications*, 125, 28-41.
- Kandhway, K., & Kuri, J. (2014). How to run a campaign: Optimal control of SIS and SIR information epidemics. *Applied Mathematics and Computation*, 231, 79-92.
- Kumar, S., Saini, M., Goel, M., & Panda, B. S. (2021). Modeling information diffusion in online social networks using a modified forest-fire model. *Journal of Intelligent Information Systems*, 56, 355-377.
- Li, C., & Ma, Z. (2022). Dynamics analysis and optimal control for a delayed rumor-spreading model. *Mathematics* 2022, 10, 3455.
- Liu, Y., Wang, B., Wu, B., Shang, S., Zhang, Y., & Shi, C. (2016). Characterizing super-spreading in microblog: An epidemic-based information propagation model. *Physica A: Statistical Mechanics and its Applications*, 463, 202-218.
- Liu, Y., Zhang, P., Shi, L., & Gong, J. (2023). A survey of information dissemination model, datasets, and insight. *Mathematics* 2023, 11(17), 3707.

- Milhazes-Cunha, J., & Oliveira, L. (2023). Doctors for the truth: Echo chambers of disinformation, hate speech, and authority bias on social media. *Societies* 2023, 13(10), 226.
- Mittal, A., Gibson, N. O., & Krejci, C. C. (2019). An agent-based model of surplus food rescue using crowd-shipping. *Proceedings of the 2019 Winter Simulation Conference (WSC), National Harbor, USA*, 854-865.
- Ojha, R. P., Srivastava, P. K., Awasthi, S., Srivastava, V., Pandey, P. S., & Dwivedi, R. S. (2023). Controlling of fake information dissemination in online social networks: An epidemiological approach. *IEEE Access*, 11, 32229-32240.
- Rosli, N., Zuber, M. F. A., & Turab, A. (2024). Mathematical modelling of a rumour spreading with the attitude of adjusting mechanisms. *Sains Malaysiana*, 53(3), 719-731.
- Rosmani, A. F., Mutalib, A. A., & Sarif, S. M. (2020). The evolution of information dissemination, communication, media and technology in Malaysia. *Journal of Physics Conference Series*, 1529(2).
- Sahnoune, M. S., Akdim, K., Ez-Zetouni, A., & Zahid, M. (2021). A virus dynamics model for information diffusion in online social networks. *Communications in Mathematical Biology and Neuroscience*, 2021(2021).
- Schreiner, M. & Riedl R. (2019, May 24-27). *Identification of factors influencing the engagement effect in social media communication* [Paper presentation]. 48th Annual European Marketing Academy Conference, Hamburg.

- Velten, K. (2009). *Mathematical modeling and simulation: Introduction for scientists and engineers*. Wiley-VCH.
- Wang, C., Wang, G., Luo, X., & Li, H. (2019). Modeling rumor propagation and mitigation across multiple social networks. *Physica A: Statistical Mechanics and its Applications*, 535, 122240.
- Wang, J., Jiang, H., Ma, T., & Hu, C. (2019). Global dynamics of the multi-lingual SIR rumor spreading model with cross-transmitted mechanism. *Chaos, Solitons and Fractals*, 126, 148-157.
- Wen, S., Haghighi, M. S., Chen, C., Xiang, Y., Zhou, W., & Jia, W. (2015). A sword with two edges: Propagation studies on both positive and negative information in online social networks. *IEEE Transactions on Computers*, 64(3), 640-653.
- Xia, J., Yu, L., & Wang, Y. (2023). A method for simulating the propagation of network hot topics based on combined SEIRS-ARIMA model. *Academic Journal of Mathematical Sciences*, 4(2).
- Yang, B., Zhang, R., Cheng, X., & Zhao, C. (2023). Exploring information dissemination effect on social media: An empirical investigation. *Personal and Ubiquitous Computing*, 27, 1469-1482.
- Zhang, Y., & Chen, Z. (2019). SETQR propagation model for social networks. *IEEE Access*, 7, 127533-127543.

APPENDICES

Appendix A: Project Schedule with Gantt Chart

Task Description	Start	End	Duration	Date	7/10/24	14/10/24	21/10/24	28/10/24	4/11/24	11/11/24	18/11/24	25/11/24	2/12/24	9/12/24	16/12/24	23/12/24	30/12/24	6/1/25	13/1/25
				wk	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Phase 1: Definitions	7/10/24	21/11/24	46																
Identify problem	7/10/24	13/10/24	7																
Establish goals, objectives, and requirements	14/10/24	19/10/24	6																
Submit brief proposal	20/10/24	20/10/24	1																
Revise proposal	20/10/24	14/11/24	26																
Submit full proposal	14/11/24	14/11/24	1																
Write and revise introduction of the project	14/11/24	21/11/24	8																
Submit Chapter 1	21/11/24	21/11/24	1																
Phase 2: Systems Analysis	21/11/24	13/12/24	23																
Scrap X data	21/11/24	24/11/24	4																
Preprocess X data	25/11/24	1/12/24	7																
Review existing literature	21/11/24	28/11/24	8																
Identify components and relationship of system	21/11/24	28/11/24	8																
Write literature review	29/11/24	5/12/24	7																
Revise literature review	6/12/24	12/12/24	7																
Submit Chapter 2	13/12/24	13/12/24	1																
Phase 3: Modelling	13/12/24	5/1/25	24																

[illegible][illegible]

