

**A NOVEL APPROACH TO DETECT FAKE NEWS USING DATA FROM
GOOGLE SEARCH SPECIFICALLY ON RECENT AND POPULAR TOPICS**

BY

PANG HUEY JING

SUPERVISED

BY

TS DR. OOI BOON YAIK

A REPORT

SUBMITTED TO

Universiti Tunku Abdul Rahman

in partial fulfilment of the requirements

for the degree of

BACHELOR OF COMPUTER SCIENCE (HONOURS)

**Faculty of Information and Communication Technology
(Kampar Campus)**

JAN 2021

UNIVERSITI TUNKU ABDUL RAHMAN

REPORT STATUS DECLARATION FORM

Title: A NOVEL APPROACH TO DETECT FAKE NEWS USING
DATA FROM GOOGLE SEARCH SPECIFICALLY ON RECENT
AND POPULAR TOPICS

Academic Session: JAN 2021

I _____ PANG HUEY JING _____
(CAPITAL LETTER)

declare that I allow this Final Year Project Report to be kept in
Universiti Tunku Abdul Rahman Library subject to the regulations as follows:

1. The dissertation is a property of the Library.
2. The Library is allowed to make copies of this dissertation for academic purposes.

Verified by,



(Author's signature)



(Supervisor's signature)

Address:

17A, Persiaran Halaman Ampang 12,
Halaman Ampang Mewah,
31400 Ipoh Perak.

Ts Dr Ooi Boon Yaik

Date: 16 April 2021

Date: 16 April 2021

**A NOVEL APPROACH TO DETECT FAKE NEWS USING DATA FROM
GOOGLE SEARCH SPECIFICALLY ON RECENT AND POPULAR TOPICS**

BY

PANG HUEY JING

SUPERVISED

BY

TS DR. OOI BOON YAIK

A REPORT

SUBMITTED TO

Universiti Tunku Abdul Rahman

in partial fulfilment of the requirements

for the degree of

BACHELOR OF COMPUTER SCIENCE (HONOURS)

Faculty of Information and Communication Technology
(Kampar Campus)

JAN 2021

DECLARATION OF ORIGINALITY

I declare that this report entitled “A NOVEL APPROACH TO DETECT FAKE NEWS USING DATA FROM GOOGLE SEARCH SPECIFICALLY ON RECENT AND POPULAR TOPICS” is my own work except as cited in the references. The report has not been accepted for any degree and is not being submitted concurrently in candidature for any degree or other award.



Signature : _____

Name : PANG HUEY JING

Date : 16 April 2021

ACKNOWLEDGEMENTS

I would like to express my sincere thanks and appreciation to my supervisors, Dr. Ooi Boon Yaik and Dr. Pradeep A/L Isawasan who has given me opportunity and enabling me to gain more knowledge in the domain of my research. Both supervisors had been providing me resourceful skills and guidance according to their experiences and knowledge in Data Science field which greatly helped me along my final year project. With their helps, I was able to finish my final year project and learned a lot of skills from them that I couldn't learn from books. I am glad that I have the opportunity working with Dr Ooi and Dr Pradeep.

Lastly, I wanted to thank my friends and family who always enlighten me throughout the process. These unconditional supports, loves and patience make me feel stronger and withstand all the stress or hard time. Thank you to all these people once again.

ABSTRACT

From the beginning of the coronavirus pandemic, the internet has become an important source of health information to the public worldwide. Anyhow, there have been widespread concerns that the novel coronavirus had caused a pandemic search for information with broad dissemination of false or misleading health information across social media. Therefore, the fact that all the online information being published is subjective to be clean and trustable are denied. Social media platforms are meant to share information and the speed of spreading fake news is unpredictable. Hence, a novel approach is used to detect fake news using data scraped from Google search specifically on recent and popular topics. This research model associating with few criteria identified throughout the research are viewed as the methods or steps on how a human being classify the real and fake news in their life. Therefore, by utilizing the criteria which are checking the source, date dispersion of articles, and the accuracy of search result, this research model can act as a current issue-related news checker that allows the public to filter out fake news published across the internet. The data is utilized in this research are scaped from Google, a search engine that allows the public to get worldwide information. Anyhow, this research will be specifically focusing on recent and popular topics for example the news regarding covid-19 that threaten the world recently. In consequence, different searching queries related to the recent and popular topics are used to scrape results from the Google search engine. The motivation behind this paper is to evaluate the criteria that could help in classifying fake news spreading across the internet. With the help of ensemble learning and the three criteria which are the number of articles from trustable website, average date difference between articles, and the average similarity score between quires and articles title, an accuracy of 73% could be obtained on the testing data and 32% on data with noise.

TABLE OF CONTENTS

TITLE PAGE	i
DECLARATION OF ORIGINALITY	ii
ACKNOWLEDGEMENTS	iii
ABSTRACT	iv
TABLE OF CONTENTS	v
LIST OF FIGURES	viii
LIST OF SYMBOLS	ix
LIST OF ABBREVIATIONS	x
CHAPTER 1: INTRODUCTION	1
1.1 Problem Statement and Motivation	1
1.2 Project Scope	2
1.3 Project Objectives	3
1.4 Impact, significance and contribution	3
1.5 Proposed Approach	4
1.6 Highlight of achievements	4
1.7 Report Organization	5
CHAPTER 2: LITERATURE REVIEW	6
2.1 Background Information	6
2.2 Structural Trend Analysis for Online Social Networks	9
2.2.1 Correlated Trend Detection	9
2.2.2 Uncorrelated Trend Detection	11
2.3 Surveys on Information Diffusion Across the Online Social Network	11
2.3.1 Detecting Popular Topics	11
2.3.1.1 The Used of Temporal and Social Terms Evaluation in Detecting Rising Topics	11
2.3.2 Modelling Information Diffusion	12
2.3.2.1 Deriving Networks of Diffusion and Influence	13
2.3.2.2 Revealing the Temporal Dynamics of Diffusion Networks	14
2.3.2.3 Formation and Dynamics of the Information Pathways	15
2.4 Multilayer Naïve Bayes Model	15
2.5 Misinformation Dissemination Trends Across the Social Media	17
2.6 ID3, Iterative Dichotomiser 3 – A Decision Tree Learning Algorithm	19
2.7 Ensemble Learning Method	20

2.8 Random Forest – A Ensemble Learning Method	21
CHAPTER 3: METHODOLOGY	22
3.1 Cross Industry Standard Process for Data Mining Methodology	22
3.2 Business Understanding	24
3.3 Data Understanding	25
3.3.1 Type of Fake News	25
3.3.2 Data Creation and Collection	26
3.3.3 Credible Sources Collection	28
3.4 Data Preparation	29
3.5 Modelling	31
3.5.1 Operationalization of Criteria	31
3.5.2 Utility Function	32
3.5.3 ID3, Iteration Dichotomiser 3 Model	33
3.5.4 Ensemble Learning Model	34
3.5.5 Random Forest Model	34
3.6 Evaluation	35
3.7 Deployment	35
3.8 Tools and Technologies Used	35
3.9 Implementation Issues and Challenges	36
3.10 Timeline	38
CHAPTER 4: EXPERIMENTAL RESULTS	39
4.1 Sample Data	39
4.1.1 Training Data	39
4.1.2 Testing Data	39
4.1.3 Competitors Data	40
4.1.4 Trustable Website sources	41
4.2 Utility Function and Models Implementation	41
4.2.1 Performance Evaluation	44
CHAPTER 5: CONCLUSION	46
5.1 Project Review	46
5.2 Future Work	47
Bibliography	48
APPENDIX A: Poster	A-1
APPENDIX B: Final year project weekly report	B-1
APPENDIX C: Plagiarism Check Result	C-1

LIST OF TABLES

TABLE	TITLE	PAGE
Table 2.3.2.1	Explanatory models with respect to the nature of the underlying network	13
Table 2.5.1	Facebook Actions	18
Table 2.5.2	Twitter Actions	19
Table 3.3.1.1	Type of fake news and respective description.	26
Table 3.5.1.1	Description of criteria.	31
Table 3.8.1	Python libraries version.	36
Table 3.8.2	Google Collab Specifications.	36
Table 4.2.1.1	Performance evaluation of testing data.	44
Table 4.2.1.2	Performance evaluation of competitor's data.	45

LIST OF FIGURES

FIGURE	TITLE	PAGE
Figure 2.2.1	CDF of ranking of topics (Political Hashtag Rankings)	9
Figure 2.2.1.1	General Influence Spread	10
Figure 2.3.2.2.1	NETRATE's normalized mean absolute Error (MAE)	14
Figure 2.4.1	The framework of Multiplayer Naïve Bayes Retweeting Sentiment Tendency	16
Figure 3.1.1	Four level breakdowns of the CRISP-DM Methodology for Data Mining	22
Figure 3.1.2	Phases of the CRISP-DM Process Model for Data Mining	23
Figure 3.1.3	Poll results on Data Mining Methodology conducted by <i>Data Science Project Management</i>	23
Figure 3.3.2.1	A sample of 10 rows of training queries regarding health domain issues with fake and real labelling.	27
Figure 3.3.2.2	A sample of 10 rows of testing queries regarding health domain issues with fake and real labelling.	27
Figure 3.3.2.3	A sample of 10 rows of competitor's data with different categories and '1' (real) and '0' (fake) labelling.	28
Figure 3.3.3.1	A sample of 10 rows of credible website links were collected across the internet.	29
Figure 3.4.1	Screenshot of competitor's data after cleaning and transforming.	30
Figure 3.4.2	Screenshot of trustable source list after cleaning.	30
Figure 3.5.2.1	Utility Function.	32
Figure 3.5.3.1	Architecture diagram for ID3.	33
Figure 3.5.4.1	Architecture diagram for Ensemble Learning.	34
Figure 3.5.5.1	Architecture diagram for Random Forest.	34
Figure 3.10	Gantt chart.	38
Figure 4.1.1.1	Sample of training dataset.	39
Figure 4.1.2.1	Sample of testing dataset.	40
Figure 4.1.3.1	Sample of competitor's dataset.	40
Figure 4.1.4.1	Sample of trustable website sources.	41
Figure 4.2.1	Table of data after executing Utility Function.	43

LIST OF SYMBOLS

$+$	Plus
$-$	Minus
\times	Multiplication Sign
\div	Division Sign
$=$	Equal
Σ	Sigma
ϵ	Epsilon Variant
\wedge	N-ary Logical And
\triangle	White Up-Pointing Triangle
\leq	Less Than or Equal To
\geq	Greater Than or Equal To
ε	Epsilon

LIST OF ABBREVIATIONS

API	Application Programming Interface
ASM	Attribute Selection Measure
CDF	Cumulative Distribution Function
CRISP-DM	Cross Industry Standard Process for Data Mining
et al.	And Others
EXP	Exponential
GUI	Graphical User Interface
ID3	Iterative Dichotomiser 3
IDE	Integrated Development Environment
IS	Information System
MACD	Moving Average Convergence Divergence
MAE	Mean Absolute Error
MCMC	Malaysian Communications and Multimedia Commission
n.d.	No Date
NETIF	Network Inference
NLP	Natural Language Processing
p./pp.	Page/Pages
POW	Power-Law
PR	Public Relation
RAY	Rayleigh
REST API	RESTful API
TSTE	Temporal and Social Terms Evaluation
Var	Variance
WHO	World Health Organization

CHAPTER 1: INTRODUCTION

1.1 Problem Statement and Motivation

The coronavirus disease pandemic has caused a huge burden to all the countries around the world and resulted more than millions of deaths. It is believed that false information are existed and are being spread uninhibitedly over the social media platforms at a noticeable speed when the public health or healthcare officials rushed to identify the virus that may spread. Misinformation can propagate across the internet without the need for any professional verification, constraints as well as any scientific proofs. (Kouzy et al., 2020). Hence, the spreading of false information in a country over social media or website could happen anytime, anywhere especially when there is a lack of information or evidence associated with a topic.

Coronavirus is a new virus and had brought chaos to the global in a blink of an eye. Therefore, there has been limited data published in the world regarding population knowledge, attitudes, and practices toward the virus. With many uncertainties, the public start to feel panic, confusion and eventually led to many undesired situations happened such as panic buying, people crowded at public transportation stations to travel back to their hometowns, etc. (Azlan et al., 2020). On the other hand, sharing of personal information of covid-19 patients or suspected to have covid-19 people are inhibited when their personal information was leaked and shared into social media (Yusof et al., 2020). False information may be shared but those exposed “patients” are becoming the target of hate and discrimination. This can be said that the citizens have big misconceptions about the virus, and they do not expect this virus to have such great devastating effects since nobody had experienced this pandemic before.

The motive of this research is to detect fake news using a method that people usually do for checking the authenticity of piece of news rather than using fact-checker. This is because fact-checker has limitations on verifying recent, popular and unchecked topics before a process of verification done that requires authority’s data from governments or any authoritarian parties. This increases the difficulties especially when government is lacking resources in collecting these unrevealing topics. Therefore, people who wanted to reveal the truthfulness of a news will normally check on the source, looking for the published date and also finding the best result that matched to their search queries. These steps have come to mind where normally will do when they are looking

for verification. Hence a novel approach has been brought into in this research to evaluate the effectiveness of the steps used on classifying the possible fake news that are being share tremendously on online social platforms. Other than that, using different modelling techniques associated with the criteria will be compared to figure out the best model that could give high accuracy in detecting a piece of fake news. Experiments on this approach will be carried out whether this is effective or useful in the future. Hereby, this could let the public have the ability to check on suspected news especially for those who do not know how to differential real or fake news and not blindly trusting the content which may twisted by the publishers.

1.2 Project Scope

The diffusion of fake news in social media had made people began to feel anxious, confused, and scared as they do not have the ability to distinguish whether which are correct, which are wrong, or even which to follow. Since there was a long history of fake news existed and it had led chaos to the society, it is important to seek solutions or approach to counter these problems so that nobody is harmed in the future. Hence, the output of this research will be a novel approach to detect fake news using data from Google search. The input created will be focusing on recent and popular topics especially in the health domain which mainly prioritizing covid-19 related news. The main focus of the project is to evaluate the relationship between the source, data dispersion, and similarity score of the queries to the articles' title in estimating the authenticity of the news and the analysis of the collected data from Google.

In order to achieve that, data sources or datasets are required to be gathered and analysed. On that account, the data sources will be collected through Google. Google is a popular search engine where users can search desire information published across the world including webpages, images, videos and many other features that could help the users find exactly what they are looking for. Hence, by using this popular search engine, data related to recently news can be scaped easily. Besides that, fake news related to coronavirus that are in English languages will be prioritised in this research. This is because majority of the news or data collected from Google are usually published either in English languages. Therefore, English news will be more considerable in this research.

1.3 Project Objectives

In this research, there are several objectives that are meant to achieve. First of all, the main objective in this research is to determine the effectiveness of the three criteria (trustable-source, date dispersion and similarity score of queries to article's title, mentioned in 1.2) on estimating the authenticity of the recently and popular news across the internet. Besides that, the project is meant to determine which of the model among ID3, Ensemble Learning and Random Forest is more suitable to be used in this research to classify the news.

That being so, in summary, the research objectives are listed below:

- To determine the effectiveness of the three criteria in this research.
- To determine the most suitable techniques implementing with the three criteria in detecting fake news.

1.4 Impact, significance and contribution

This research will help the society to have an early involvement towards false information especially to those responsible groups, organizations, or even government. This is because there is still a substantial amount of people who are not aware of the presence of manipulated images, videos, sponsored content specifically elders who are not familiar with the online information. Perhaps, teenagers or late adolescence who are actively involved in social media might also have difficulties detecting lies, have greater trust in what is spreading through the internet, or eventually ignored the accuracy of information (Nadia and Daniel, n.d.).

Hence, this research is delivering a novel approach on how to detect fake news that surrounded the society every single day. There are many news or articles published through the internet were seen to have a higher chance of spreading issues where the terms or hashtags, images, or videos are commonly used in articles or posts. However, looking at these features that are likely associated with misinformation only are not enough. Therefore, there are three criteria are identified to verify the authenticity of recent and popular topics searching from Google. These three criteria are the trustable publishers, date dispersion of the news and the similarity of a query compare to the article's title. This idea came to mind as some of the people will first check the source of the news whether there are from authorized parties or publishers, next, the date of

news released, and the last whether the articles or results returned by Google search engine match what they are looking for. Thus, these criteria are believed to be useful in detecting fake news especially in recent and popular topics across the internet.

1.5 Proposed Approach

This research was initiated by creating a training data for model training purposes where the data comprised of a list of queries with the labels real or fake. These real and fake queries were created according to the current issues and recent popular topics which were the health-related news especially the coronavirus pandemic that threatens the society. Besides that, testing data were also created by implementing the criteria of recent popular topics in health-related news which are similar to the training dataset. This testing data are mean for the testing purpose as it allowed us to evaluate the accuracy of the research approach on detecting fake news. At the same time, a set of competitors' data was also created by extracting the Fake News data from GitHub.

A utility function was created where this function is a mathematical formulation that grades the inclinations of the person regarding their satisfaction in choosing or consuming different bundles provided. This utility function was implemented in Google Collab using Python default versions Py 3.6.7. This mathematical formulation is meant to calculate the three criteria stated before, as each of these criteria will return a list of numbers regarding the dataset used. In this situation, the effectiveness of criteria can be analysed accordingly to the rank of preferences to each of the criteria.

Information extracted from the raw data are intercepted before it was transfer to the 3 models founded on Decision tree learning using ID3 (Iterative Dichotomiser 3), Ensemble learning, and Random Forest. A set of results will be generated using these algorithms from each of the datasets where the datasets are training, testing, and competitors' data. As consequence, an overview can be obtained by comparing the results from these datasets and equivalent to saying that the objectives can be achieved.

1.6 Highlight of achievements

An accuracy of more than 70% on 30 testing data could be achieved by the model established on Ensemble learning method while the accuracy of the ID3 model came second with a 70% accuracy and the Random Forest model came last with an accuracy

of 50% which barely achieved the benchmark accuracy score. However, when the data used for testing the models are not from a similar scope, the models could only predict one-third of them correctly. This is expected as the models expect covid-19 related news and some other popular or current news which focused on health domain instead of news from sport domain or from a political domain.

1.7 Report Organization

The details of this research are shown in the following chapters which included a total of 5 chapters. The first chapter is a brief introduction that delivered the research background information, problem statement, motivation, objectives, scope, impact, and contribution. For the second chapter, several existing research and papers regarding this research background are being reviewed and studied in this chapter. In chapter 3, the methodology used for the proposed method or approach is described while experiments and results that are obtained for this research will be covered in chapter 4. Lastly, chapter 5 will give a conclusion on this research.

CHAPTER 2: LITERATURE REVIEW

2.1 Background Information

In recent years, online platforms are widely used by the public for a variety of purposes. Nowadays, using social media platforms to share online contents, messaging, interacting with friends or family is a norm among the people. Moreover, online platforms had brought benefits to the people as they are much easier to access, lower cost, and comes in rapid dissemination of information. People choose to consume more news through the internet rather than with traditional news media, such as newspapers, radio, or television. According to About Facebook (2019), there are over 3 billion people around the world who own at least one Facebook account. While for Twitter, there are around 330 million monthly active users and around 140 million daily active users who are retweeting and viewing pieces of stuff on Twitter (Ying, 2019). This can be said that social media have experienced aggressive growth in user base as compared to the past few years. Moreover, online platforms were later adopted by business companies who wish to take advantages of a popular new communication method to reach out to customers. This enables the widespread of misleading information which is also known as fake news. Fake news is the news with false information intentionally. Consequently, the potential for negative impacts on individuals or society had given a sharp warning of red light with the extensively spread of misinformation.

In the political domain, online platforms are also being viewed as one of the factors that will affect the result. Although there is a lengthy history of misinformation and mass misinterpretations on the political process, the online platforms had shown as an ideal platform for the politicians, political parties, political foundations, etc to spread their opinions publicly through the networks widely in recent years. These actions are trusted to have the ability to increase political participation (Bruns and Highfield cited in Stieglitz et al., 2014). For example, the most popular fake news had a higher exposure rate on social media than the most popular mainstream news during the 2016 election (Silverman cited in Allcot and Gentzkow, 2017). Above 50% of US, adults get news through the internet and the majority of them who read that fake news were reported that they believe them (Gottfried and Shearer cited in Allcot and Gentzkow, 2017; Silverman and Singer-Vine cited in Allcot and Gentzkow, 2017). For this reason, many

argue that false news played a vital role in the 2016 election as stated in Parkinson, 2016; Hunt, Matthew, and Chuan, 2018.

In addition, the growing assumption of online platforms has brought a great potential in increasing the interactions between citizens and politicians in Australia at least since the 2007 federal election in Flew, 2008; Kirchoff et al., 2009 (Bruns and Highfield, n.d.). The level of participation in public debate is being raised as the citizens are able to interact with the politician through online platforms like Facebook, Twitter, etc. Thus, politicians see these platforms as an opportunity to connect with their supporters as it presents a chance of making improper statements utilized by the political opponent. These styles of activities are always practiced by those politicians who are assured of winning the electoral contest. On the contrary, those politicians who are more likely to be defeated during the elections may choose to exploit social media as a last line of defence to congregate their voters or supporters and campaign strenuously on social media platforms. As such, those candidates who locked in a tight election seem to be more likely to use this medium to engross and challenge their opponents, provoke the other side into anger or hoping to win the debate where all these are publicly visible and are higher potentially accessible by all the users on the social media platforms (Bruns and Highfield, 2013). This can be said that social media have become an ideal platform for politicians on developing their strategies internationally to win the contest and this has been verified as a trend since lately years.

For another example, during the coronavirus pandemic, the internet has become a major source of health information to the public worldwide. For instance, the novel coronavirus caused a massive search for information with many misleading health information across the network. Therefore, the fact that all the online information being posted is subjective to clean and trustable are undeniable. On 15 February 2020, the WHO Director-General Tedros Adhanom Ghebreyesus said that the infodemic has endangered the society. Fake news spreads faster and more effortlessly than a virus, and it just as dangerous” during the conference. The unpredictability surrounding the coronavirus, paired with the intense global demand for information had created a difficult situation of prediction, deception, and spreading of false or even harmful information.

Besides that, according to Sylvie Briand, Director of Infections Hazards Management at WHO every outbreak will be accompanied by the enormous amount of information with misinformation and rumours in it. With the presence of social media platforms, the phenomenon is further amplified, and it will be a challenge to the people during the outbreak. Even though many states across the countries had implemented the stay-at-home orders to beat up the pandemic, but the divergence of public health conditions, lacking of confidence on beating up the pandemic is more likely to lead to the sustained growth of misinformation and eventually causing chaos in the society.

Based on the last few paragraphs, misinformation had truly driven the crisis across the world. Malicious actors took advantages of confusing, fear, and sorrow online users for profit and political gain by deliberately spreading misrepresentation, tricks, and stoking engagement among the community. This seems to be a gift to them in this modern technologies' environment especially during this pandemic where the people seek for the latest news or information. Citizens around the world who are surrounded by the fake news are in danger and eventually, there will at least a few hundreds of people who might die because of the misinformation. What's more, is that the older generation who are not familiar with the cyber world are having the highest risks among the online platform community as they are easily influenced by the news in the cyber world. All of the users across the online platforms are also having difficulties in distinguishing the facts and fake news as fake news spread faster and more easily. Ultimately, those articles that were written or posted with widely shared counts and more likes were more likely to grab users' intentions and eventually, make the users believe that they are true.

Based on Welle (2020), there are around 800 people were died from drinking highly concentrated alcohol to disinfect themselves in the hope of fighting against the virus, while around 5900 citizens who consumed methanol were hospitalized and 60 people went blind. Not only that, but there are also are a lot of conspiracy theories and rumours that were posted in the social media such as drinking cow urine, eat dung, camel urine with lime, eating garlic, etc to prevent infection. Even though all these nonsenses are not scientifically proven by the scientists, all these news are still being widely spread and believed by the citizens that such actions could prevent themselves from getting infected by the virus. From this point of view, it is important to confirm the integrity of the news before sharing and the awareness of every citizen across the globe should be

increased in order to prevent those problems that are stated above from happening continually.

2.2 Structural Trend Analysis for Online Social Networks

In this paper, the authors had given some engrossing insights on how people share information using Twitter. Over 500 hashtags were being analysed by the authors. These hashtags are classified into 7 different themes such as technology, celebrity, idioms, movies, political, games, and music. By referring the Figure 2.2.1, Cumulative Distribution Function, CDF is implemented on the ranking of topics of the political category under correlated, uncorrelated and traditional trends. The importance of political hashtags is reinforced using the correlation trends definition. This had indicated that political tags have high correlated importance where it reveals that people will share the information that had been shared by their friends or simply that these people have similar political views. In the meantime, the authors also provided some efficient methods for both correlated and uncorrelated trend detections as their goal is to give a ranked list of top-k topics for both trend definitions.

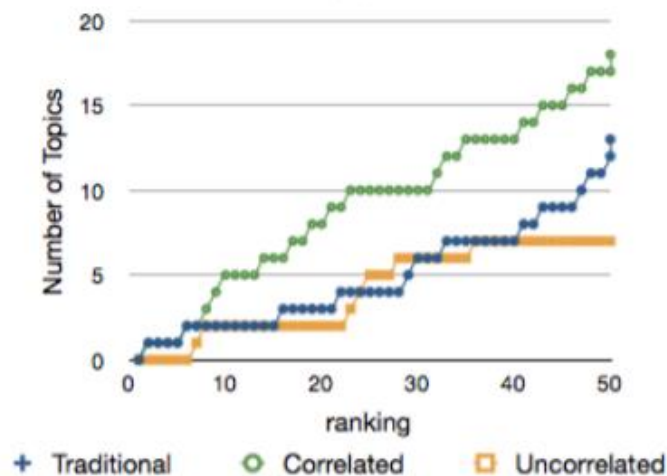


Figure 2.2.1 CDF of ranking of topics (Political Hashtag Rankings)

2.2.1 Correlated Trend Detection

In this section, a straightforward sampling method is used by the author while the high accuracy is still guaranteed at the same time. The problems of assessing the significance of each topic with regard to the correlated trendiness view to a problem of counting local triangles are lessened in order to demonstrate the use of the sampling method. Counting the number of triangles incidents at a given node in a graph G shown as an

example of this technique. The authors proposed their sampling-based solution to the extensibility challenge of correlated trend detection. To ease the correctness of the solution in a graph-oriented manner, the authors displayed the problem of finding correlated trends is correspondent to counting local triangles in a multigraph.

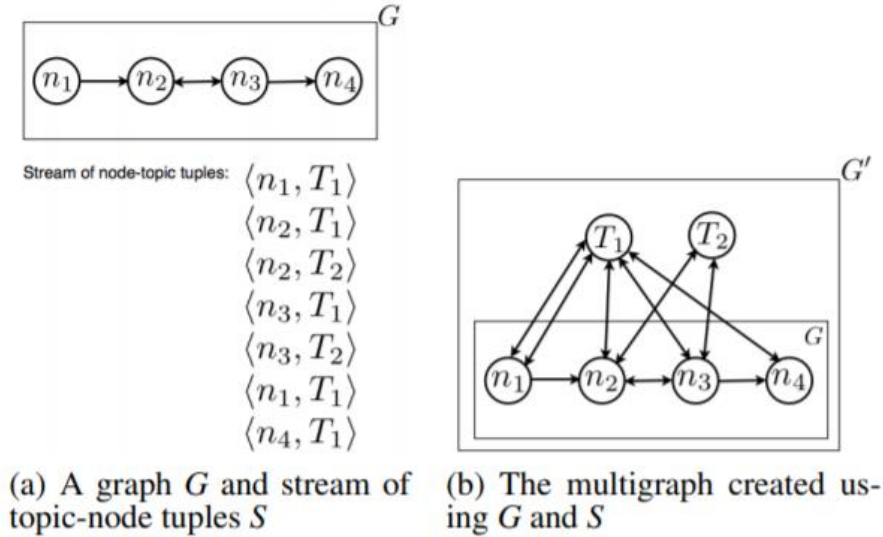


Figure 2.2.1.1 General Influence Spread

At the same time, the authors wished to assurance the number of triangles calculated is best estimated of the total number of triangles. Particularly, the number of triangles involving T_x in G' can be considered as $X_x = Count_x / p_s^2$ and the probability of the prediction X_x is off by $\epsilon \Delta_x$ is upper bounded by the following equation:

$$Pr(|X_x - \Delta_x| \geq \epsilon \Delta_x) \leq \frac{Var(X_x)}{\epsilon^2 \Delta_x^2} \leq \frac{(p_s^2 - p_s^4)}{p_s^4 \epsilon^2 \Delta_x} + 2\alpha_x \frac{(p_s^3 - p_s^4)}{p_s^4 \epsilon^2 \Delta_x^2}$$

Implementation of the equation above could prove the standard of the estimation depends on the number of the triangles and the number of edge-disjoint triangles. Because the number of multi-edges brought a huge impact on this property, the standard of estimation depends on the number of times a specific topic is mentioned by a particular user. When this number escalated in a larger range, the quality of the estimation will also be affected, hence decreases. This shows that it is better for trendy topics even though the estimation gone linearly and worse with α_x , but is still equilaterally better with increasing Δ_x .

2.2.2 Uncorrelated Trend Detection

For uncorrelated trend detection, it is mostly similar to correlated trends. In a multi-graph, counting local triangles could be achieved by reducing uncorrelated trend detection. Similar to what the author proposed in 2.1.1, the problem can be efficiently estimated with sampling. The uncorrelated trendiness score of topics should be gradually updated as an online algorithm is needed. The exact increment can be calculated in the following way:

$$h'(T_x) = h(T_x) + \sum_{n_i \in (N - n_l - N_l)} C_{i,x} + \sum_{n_i \in (N - n_l - N_l')} C_{i,x}$$

During each update operation, it uses the power-law degree distribution of social networks with a small-scale of reads. Since the huge number of triangles are lower than the quality of sampling and the thinness of social network graphs, the result for uncorrelated trends are similar to correlated trends, but it is way more robust on sampling.

2.3 Surveys on Information Diffusion Across the Online Social Network

2.3.1 Detecting Popular Topics

Detecting popular topics is one of the major tasks to develop automatic means to give an overview of the topics that are being favoured by the public or will be a popular topic in the future. Topic detection techniques with a traditional way to analyse static corpora are no longer alter to message streams nowadays. Kleinberg cited in Guille et al. (2013) had proposed a state machine to examine the arrival times of stream documents in order to distinguish bursts as documents will have a similar theme. A bursty topic is defined as a topic with a behaviour that has been suddenly treated within a time or interval.

2.3.1.1 The Used of Temporal and Social Terms Evaluation in Detecting Rising Topics

The authors had proposed a topic detection approach where it is known as Temporal and Social Terms Evaluation (TSTE). This method analyses both the temporal and social properties of the diffusion of the messages. In the beginning, these authors had

developed a process that consists of five steps where the first step is to utilize the augmented normalized term frequency. To turn the message's content into vectors of terms with comparative frequencies computed. Next, the authority of the dynamic creators is evaluated using their relationship and the Page Rank algorithm. According to the paper entitled written by Page et al. (1999), the authors proposed this PageRank as it is a mechanism for computing a ranking for every web page that regarding the graph of the web. Besides that, it also permits to display the existence pattern of each term based on a biological metaphor, which depends on the value calculated for sustenance and vitality that influence the users' authority. The proposed method can identify most of the bursty terms by using supervised or unsupervised algorithms, as the main purpose of calculating a critical drop value is based on the energy. Lastly, bursty topics are derived into sets of terms with the help of co-occurrence-based measurement.

2.3.2 Modelling Information Diffusion

In this section, analysis on how misinformation spread will be studied more deeply. Some models are suggested to seize or predict the spreading process in online social networks. The diffusion process is categorized into two which are activation sequence and spreading cascade. The activation sequence is an ordered set of nodes that captured the order where the nodes of the network acquired information whereas a spreading cascade is meant to a directed tree that has a root as the first node of the activation sequence. This tree will capture the influences between the nodes and unfolds in the same order that adopted in the activation sequence. The authors summarized the surveyed model in Table 2.3.2.1 and each of the papers will be reviewed according to the original papers.

reference	network		inferred properties			supports missing data
	static	dynamic	pairwise transmission probability	pairwise transmission rate	cascade properties	
<i>NETINF</i>	x		x		x	
<i>NETRATE</i>	x		x	x	x	
<i>INFOPATH</i>	x	x	x	x	x	

Table 2.3.2.1 Explanatory models with respect to the nature of the underlying network.

2.3.2.1 Deriving Networks of Diffusion and Influence

Correlation in nodes infections times that infer the structure of the spreading cascade is explored and proposed by the authors. They assume that the activated nodes had affected each of their surrounding nodes with some probability. They come out with an approach which is NETIF, which is a scalable algorithm based on submodular functions to find the spreading cascade for deriving networks diffusion and influence. The generative probabilistic model is created to study on how the viruses are spread through the network initially. There are numerous potential ways the infection could take part through the network that are stable with observed data since they only observed times when the nodes get infected. Hence, in order to deduce the network, they had thought about all the possible ways the viruses could spread through the network. Subsequently, credulous calculation of the model takes exponential time slice as there is a large number of propagation trees. Surprisingly, the calculation over the set of trees can be executed in polynomial time. But a problem existed wherewith such model, the network inference issue is still recalcitrant. Therefore, the authors present a manageable estimation, to show that the target capacity can be both effectively figured and productively upgraded. Although NETIF has its own strengths that allow us to have more ideas on how the correlation in node infections times and the deduce of spreading

cascade, but it also has its weaknesses where it cannot support dynamic network, pairwise transmission rate in inferred properties and it also does not support missing data during the analysis.

2.3.2.2 Revealing the Temporal Dynamics of Diffusion Networks

As there exists a space for improvement for the NETIF approach, Gomez et al. extend NETIF by adopting the weakness in the previous paper and propose a model that could change the diffusion process into a continuous spatially discrete network that is independent to the temporal processes at different rates. A probability density function is used to display the probability of node infecting another node upon the infection times and transmission rate in two nodes at a guaranteed time. In order to formulate and fix the pairwise transmission rates and the graph of diffusion that were raised in most of the extreme probability issues, NETRATE is proposed. In addition, the accuracy of NETRATE has been discussed in this paper too.

False edges or no edges inferred networks will only have zero accuracy. Inferencing accuracy of the NETRATE on transmission rates over edges by calculating normalized mean absolute error (MAE) is investigated. Estimation of transmission rates for the three types of Kronecker networks and a Forest Fire network is shown, which is computed on over 5000 cascades using normalized MAE as shown in Figure 2.3.2.2.1. This result had considered all three models of transmission.

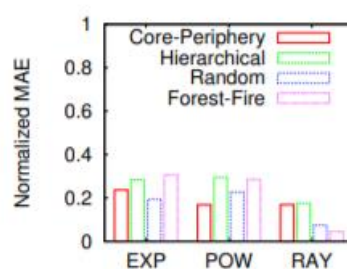


Figure 2.3.2.2.1 NETRATE's normalized mean absolute Error (MAE)

While in this paper, the NETRATE performance related to cascade coverage and time horizon is also evaluated by the authors. A more accurate estimation of the transmission rates and a higher precision-recall value could be achieved by having more cascades. Since the estimation of transmission rate is tough the authors require more cascades for

accurate estimates. Besides that, as the duration of the observation window increases, the NETRATE's accuracy will be increased over the time and deduce transmission rates. Anyhow, this method considers that the fundamental network remains static over time. Therefore, it also consisted of some weaknesses that needed to improve too.

2.3.2.3 Formation and Dynamics of the Information Pathways

According to Guille et al. (2013), the implementation of NETRATE is not satisfying for assumption. This is because the topology of online social networks is evolving very fast in terms of edges creation and deletion. Consequently, Gomez et al. extend their previous approach – NETRATE to INFOPATH. A time-varying inference algorithm, INOPATH which uses hypothetical gradients to estimate the structure and the periodically changing temporal dynamics of the network is introduced in this section. Since all of the network inferencing algorithms above are considered as static, the authors had to assume that the information propagates statically over the time. Since the initial idea of the proposed approach was to understand the dynamic edge transmission rate and how the information pathways fade in and fade out over the time, the authors had no choice but to run the method. Information route over general repetitive themes spread is proved remain constant or stable over the time with the help of the proposed method which took in real data as input. Conversely, major real-world occasions lead to emotional changes and moves in the information routes. Anyhow, their work will open different scenes for future work as it has fulfilled major strengths shown in Table 2.3.2.1 where NETIF and NETRATE do not.

2.4 Multilayer Naïve Bayes Model

Naïve Bayesian classifier is one of the supervised text classifiers that take in statistical algorithms from machine learning to be a standard method for automated text mining (Stieglitz et al., 2014). According to Bermingham and Smeaton cited in Wang et al. (2015), a group of researchers found that Naïve Bayes was a good method to classify microblogging. Wang, Zuo, and Wang (2015) mention that the analysis is defined in a given group of retweets that are related to discrete feature vectors and respective retweeting sentiment tendency. By referring to Figure 2.4.1, there are three modules which include the Naïve Bayes model on 3 of the layers. The user's profile and emotion will be the main focus of the first layer, while the user's relationship will be predicted

by the middle layer while the top layer will be focus on the user's retweeting sentiment tendency prediction.

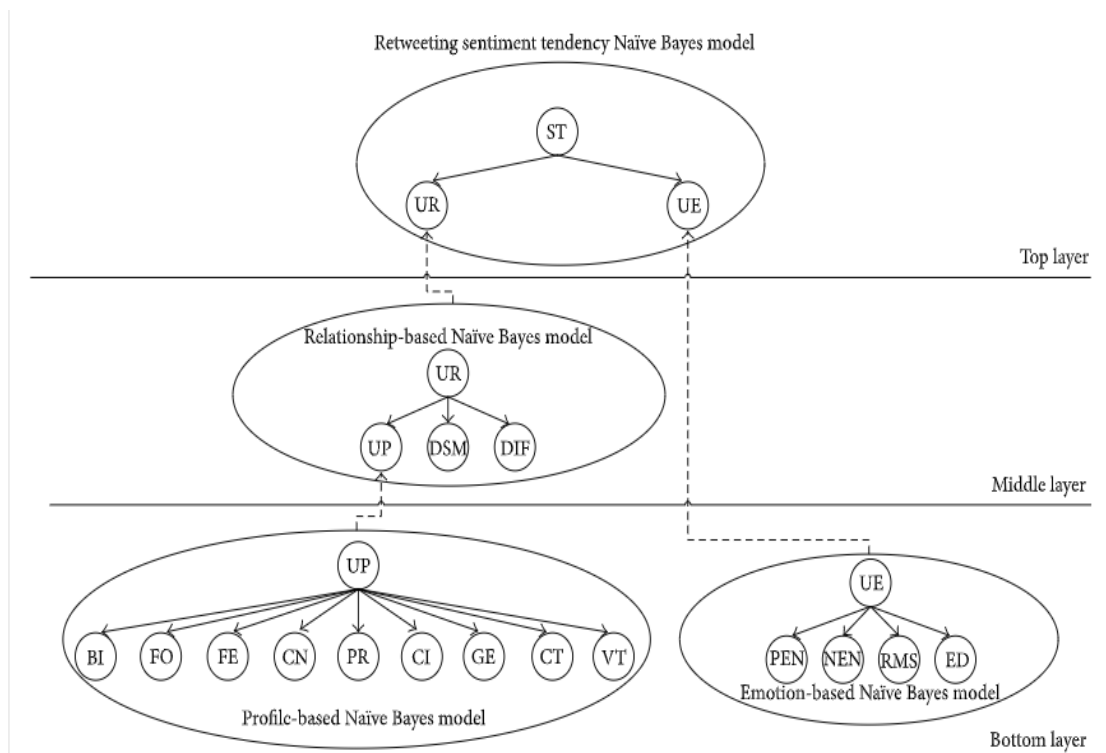


Figure 2.4.1 The framework of Multilayer Naïve Bayes Retweeting Sentiment Tendency.

Profile-based Naïve Bayes model and emotion-based Naïve Bayes model are in the bottom layer of Naïve Bayes models. However, the emotion-based Naïve Bayes will not discover in deeply as our research scope does not cover on this. Bayes' theorem is the backbone of the Naïve Bayes model. With the help of conditional independent assumption to classify unseen samples by ignoring the maximum prior probability, computation overhead could be reduced. The variables that are required in the calculation are denoted as meaningful variable names. $P(\cdot)$ denotes the probability of a feature equals to discrete value. With the help of the middle layer Naïve Bayes model, dynamic Salton metrics and the dynamic interaction frequency, the users' relationship and profile could be calculated and determined. Therefore, the user's relationship can be considered as the root node of Naïve Bayes model while the remaining item will be the leaf nodes. For the following layer – top layer, Naïve Bayes model is used to calculate the user's retweeting sentiment tendency, user's relationship, and profile.

Thus, a multilayer Naïve Bayes model is suitable to be used for analysing the user's retweeting sentiment tendency. The user's relationship, retweeting sentiment tendency will be referred as root nodes while the user's relationship will be referred as leaf nodes in this layer. In conclusion, the multilayer Naïve Bayes model could find out all the variations and analyse the user's retweeting sentiment tendency on Twitter.

2.5 Misinformation Dissemination Trends Across the Social Media

According to Allocott, H et al. (2019), new evidence regarding the volume of misinformation from 2015 to July 2018 had flowed on social media platforms. The volume of Facebook engagements and Twitter shares were measured by the authors monthly in order to utilize the data they collected from BuzzSumo (www.buzzsumo.com). Apart from Facebook and Twitter, stories from major, minor news sites that were not delivering deception and sites and culture and business-related sites were also measured. This proved that data on social media can be utilized to comprehend inquiries in political theory during media presentation and content moderation practices on social media platforms.

During the data collection phase, there is a list of sites that consist of five lists from different sources creating false information are compiled by the authors. Facebook works well with PolitiFact and FactCheck to evaluate the correctness of potentially false stories flagged by Facebook users. This list mainly focuses on most important contributors of the fake news as questionable articles could be flagged by Facebook users themselves for review. Anyhow, the lists give a total of 672 distinctive fake news sites. Besides that, there were 3 additional lists of fake sites which excluded sites that do not produce similar content were collected by the authors. The 3 additional lists along with the 5 lists collected previously were then cross-validated with the BuzzSumo database. Thus, only 739 sites were leftover in the end as sites with missing data in the BuzzSumo database were then further removed from the lists.

Date	Actions
Dec 15, 2016	Announced four updates to address fake news: make reporting easier for users; flag stories as “Disputed” in collaboration with fact-checking organizations and warn people before they share; incorporate signals of misleading articles into rankings; and disrupt financial incentives for spammers. ²
Apr 6, 2017	Described three areas where it is working to fight the spread of false news: disrupt economic incentives; build new products to curb the spread of false news; and help people make more informed decisions. ³
Apr 25, 2017	Tested “Related Articles”, an improved feature that presents users a cluster of additional articles on the same topic when they come across popular links-including potential fake news articles-to provide people easier access to additional information, including articles by third-party fact checkers. ⁴
Aug 8, 2017	Announced it would address cloaking so people see more authentic posts. ⁵
Aug 28, 2017	Announced it would block ads from pages repeatedly sharing false news. ⁶
Dec 20, 2017	Announced two changes to fight against false news: replace “Disputed” flags with “Related Articles” to give people more context; and start an initiative to better understand how people decide whether information is accurate. ⁷
Jan 11, 2018	Prioritized posts from friends and family over public content. ⁸
Jan 19, 2018	Prioritized news from publications rated as trustworthy by the community. ⁹
Jan 29, 2018	Prioritized news relevant to people’s local community. ¹⁰
May 23, 2018	Described three parts of their strategies to stop misinformation: remove accounts and content that violate community standards or ad policies; reduce the distribution of false news and inauthentic content; and inform people by giving them more context on the posts they see. ¹¹
June 14, 2018	Detailed how its fact-checking program works. ¹²
June 21, 2018	Announced five updates to fight false news: expand fact-checking programs to new countries; test fact-checking on photos and videos; use new techniques in fact-checking including identifying duplicates and using “Claim Review”; take action against repeat offenders; and improve measurement and transparency by partnering with academics. ¹³

2.Addressing Hoaxes and Fake News.

3.Working to Stop Misinformation and False News.

4.New Test With Related Articles.

5.Addressing Cloaking So People See More Authentic Posts.

6.Blocking Ads From Pages that Repeatedly Share False News.

7.Replacing Disputed Flags With Related Articles.

8.Bringing People Closer Together.

9.Helping Ensure News on Facebook Is From Trusted Sources.

10.More Local News on Facebook.

11.Hard Questions: What’s Facebook’s Strategy for Stopping False News?

12.Hard Questions: How Is Facebook’s Fact-Checking Program Working?

13.Increasing Our Efforts to Fight False News.

Table 2.5.1 Facebook Actions

By analysis the likelihood of the sites from the datasets, the authors found that the interaction of the fake news sites on both social media platforms increased steadily from the beginning of 2015 up to the 2016 election. However, the interaction of fake news sites was found reducing tremendously (more than 50%) on Facebook after the 2016

election that was still found increasing on Twitter. Besides that, the authors found that Facebook had done a great job in reducing the misleading fake sites from their platforms according to Table 2.5.1 as compared to Tweeter who only started in tackling the problems in mid-June of 2017 as shown in Table 2.5.2. Moreover, the authors also have done some checking in order to ensure the robustness of the analysis. Graphs were created based on fake news sites that had made on occurrence on multiple lists and double counting of black domains that are derived from the 3 black domains provided by (Grinberg et al., 2018) is avoided.

Date	Actions
June 14, 2017	Described the phenomenon of fake news and bots and the approaches that Twitter used, including surfacing the highest quality and most relevant content and context first, expanding the team and resources, building new tools and processes, and detecting spammy behaviors at source. ¹⁴
June 29, 2017	(Not officially announced) Tested a feature that would let users flag tweets that contain misleading, false, or harmful information. ¹⁵
Sept 28, 2017	Shared information on its knowledge about how malicious bots and misinformation networks on Twitter may have been used in the 2016 U.S. Presidential elections and its work to fight both malicious bots and misinformation. ¹⁶
Oct 24, 2017	Announced steps to dramatically increase the transparency for all ads. ¹⁷
July 11, 2018	Announced the removal of fake accounts. ¹⁸

14.Our Approach to Bots & Misinformation.

15.Twitter is looking for ways to let users flag fake news, offensive content.

16.Update: Russian Interference in 2016 US Election, Bots, & Misinformation.

17.New Transparency For Ads on Twitter.

18.Confidence in follower counts.

Table 2.5.2 Twitter Actions

2.6 ID3, Iterative Dichotomiser 3 – A Decision Tree Learning Algorithm

ID3 is an algorithm used to generate decision trees from a dataset where it could give a precise overview on interpreting the feature vectors. Decision trees are often used to gain important information by giving values for the unknown object before identifying a suitable classification based on the decision tree rules. It is easy for human-level thinking as it gives a flowchart-like structure of the tree that helps in decision making. A decision tree is a tree where it consists of nodes, branches, and leaves. The node representing a feature or attribute from the dataset, while the branch indicating a decision or rule, whereas the leaf representing the outputs of the tree. The ID3 algorithm

is usually implemented in machine learning and natural language processing (NLP) domains.

In ID3, it always follows an acquisitive approach by choosing the best attribute using attribute selection measure (ASM) in classification. Attribute Selection Measure is also known as splitting rules as it helps to identify the breakpoints on given nodes. In that case, some popular selection measures are Information Gain, Gain Ratio, and Gini Index. For ID3, it used Information Gain to find the best attributes.

Information Gain is a formula where it computes the difference between the entropy before and after splitting of the dataset regarding the given attribute values. Entropy measures the uncertainty or impurity of the dataset. A value of 0 entropy suggests that it is a pure class while entropy with value 1 indicating all are the same category. In simple words, the highest information gain of a feature will be chosen as the best attribute. Thus, an overview of steps carried out in the ID3 algorithm can be defined. It will first calculate the entropy for the learning datasets. Then it will calculate the information gain for each attribute or feature so that a decision tree node could be figured out by finding the maximum information gain. These steps will be run iteratively until the desired tree is obtained.

2.7 Ensemble Learning Method

There are many methods that could be used to construct an ensemble. There is Bayesian voting where the hypotheses are enumerated and weighted by its posterior probability, manipulation of training data through bagging or sampling, ADABOOST algorithm along a final classifier is built to weight the vote of each of the individual classifiers that weight the data based on what they were trained, manipulating the input features, manipulating the output targets and lastly by injecting randomness to the learning algorithm. These are some of the common methods used to construct an ensemble, but which ensemble method is the best among these methods.

According to Dietterich's study, ADABOOST ensemble method normally return a better result as compared to bagging and randomized tree while bagging and randomized tree are having almost the similar performance. However random tree does perform better in some cases where the data sets are large when compare to bagging.

In Dietterich's experiment, ADABOOST could perform better in low-noise cases without overfitting the ensemble however overfitting issues occurred when it is used in high-noise cases. However, Bagging and Randomization both could perform well in low noise cases and high-noise cases. A more aggressive ADABOOST is also created by Dietterich in his experiment to prove that the overfitting issues faced by the standard ADABOOST is due to the stage-wise optimization process where the ensemble would slowly overfit the data.

In conclusion, by combining a less accurate classifier with ADABOOST, an ensemble that could obtain higher accuracy on that particular classifying task could be obtained.

ADABOOST is widely used in machine learning as it helps to create a more robust model by combining the old models with lower accuracy.

2.8 Random Forest – A Ensemble Learning Method

Random forest is one of the ensembles learning methods used by many researchers for classification, regression, and other tasks that conduct by building a number of decision trees during the training phase. According to Ali, Khan, Ahmad, and Maqsood (2012), In random forest, each decision tree is trained with randomly selected or random sampled data from the original data pool. Since features are randomly selected in each decision spilt, the correlation between each tree is reduced while a more complex model or classifier is produced, When the correlation between each tree is reduced, overfitting issues could be avoided, and outliers are more acceptable to the model. A random forest classifier commonly outperforms decision trees classifier. This is because a random forest has a voting mechanism that made the final decision for the classifier and the benefits of the decision tree are retained.

A random forest is commonly used to deal with data that have more features as its ability to tolerate missing data and handling continuous, categorical and binary data. By utilizing the ensemble strategies and random sampling of data random forest could achieve better generalizations and made more accurate predictions without worrying about the overfitting issues. Besides that, when the voting system of the random forest is replaced with a bagging scheme, the random forest will be more generalized and able to identify the importance features in a dataset. The accuracy of the classifier might also be improved when this scheme is used instead of a voting scheme.

CHAPTER 3: METHODOLOGY

3.1 Cross Industry Standard Process for Data Mining Methodology

CRIPS-DM is used as the methodology for this research. According to Wirth and Hipp (2000), CRISP-DM, which stands for Cross Industry Standard Process for Data Mining is portrayed regarding a progressive cycle model, containing four degrees of abstraction. The four levels of abstraction are phases, generic tasks, specialized tasks, and process instances. These four levels have different tasks and phase will be processed accordingly, refer to figure 3.1.1.

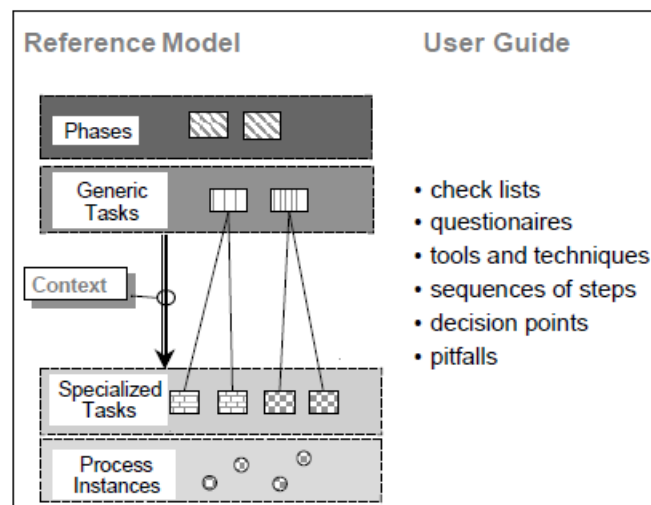


Figure 3.1.1 Four Level Breakdowns of the CRISP-DM Methodology for Data Mining

However, the main focus on the CRISP-DM Process Model in this research is the life cycle of a data mining project. The CRISP-DM reference model for information mining gives a diagram of the life cycle of a data mining project. It contains few phases required in a project which are their respective tasks and also their outputs. In the life cycle of a data mining project, there are six phases are shown in Figure 3.1.2 such as business understanding, data understanding, data preparation, modelling, evaluation and deployment.

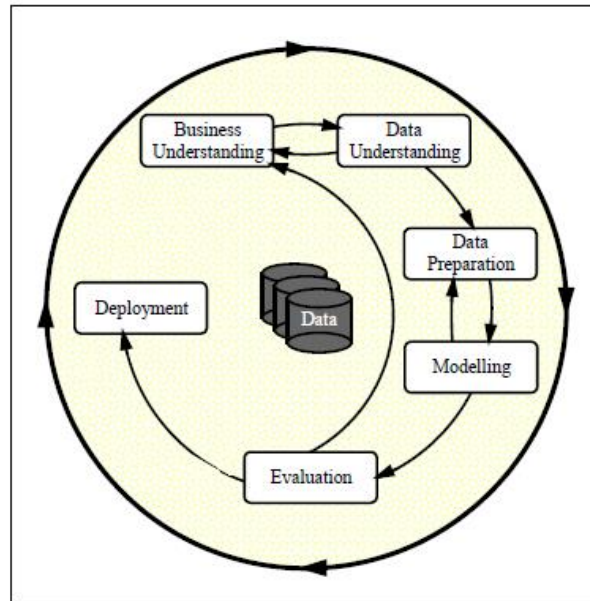


Figure 3.1.2 Phases of the CRISP-DM Process Model for Data Mining

CRISP-DM is the most often used methodology for data mining projects. This is being proven by a poll conducted by *Data Science Project Management* in August and September 2020. According to this poll, there are 109 votes for CRISP-DM as the most common use approach (Data Science Project Management, n.d.). The result of the poll conducted is as below in Figure 3.1.3. Therefore, the six phases in the life cycle of CRISP-DM will be carried out throughout this research and allow the audience to have a better understanding of each of the processes done.

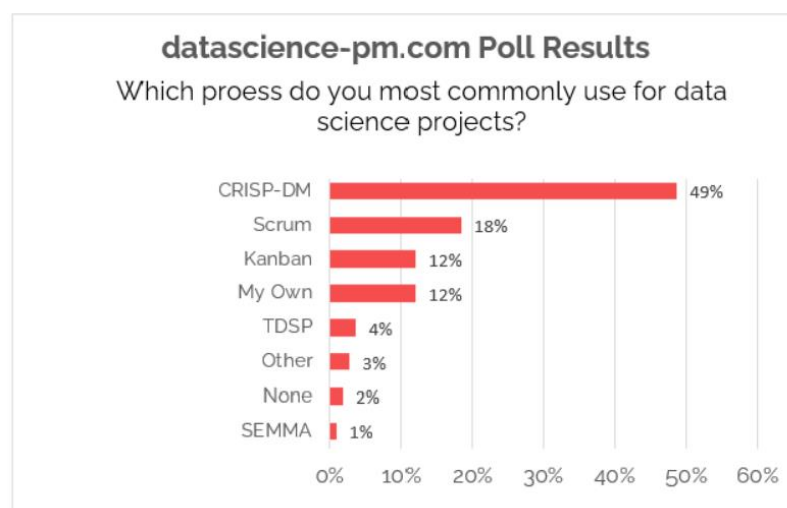


Figure 3.1.3 Poll results on Data Mining Methodology conducted by *Data Science Project Management*.

3.2 Business Understanding

This research is initiated by the business understanding phase as it is crucial to clarify the business problems. Social media and online networking have become a significant domain in information system (IS) research. A recent interest in “Big Social Data” (Manovich, 2012; Burgess and Bruns, 2012 cited in Stieglitz, 2014) had been driven partially by encouraging admittance to large-scale of datasets where these datasets are from popular social media platforms like Twitter and Facebook as well as from other online platforms such as Google that encourage mass collaboration and self-organization. This interest is not only for research but also for practical purposes. To illustrate this deeply, companies see the occasions for targeting advertising, public relations (PR), social customer relationship management, business intelligence, etc. (Stieglitz, 2014). Aside from that, political institutions also have shown interest in monitoring the public perspective on political issues, policies, political positions, managing their reputation through online platforms, etc. In addition, clickbait is another technique that is always utilized by those media companies to lure users’ attention for advertising purposes. Equally saying that more clicks on articles or advertisements mean more money will be earned by these companies.

Fortunately, there have been many developed tools exist across the internet for the user to checking the genuineness of a piece of news. For example, Politifact, FactCheck.org, Snopes, etc that can help the users to define the trustworthiness of the articles. Anyhow, these automated fact-checkers have their limitations in terms of what they can achieve by themselves. This is because an automated fact-checker is less capable in distinguishing complicated claims or sentences with the subtler structure of false information compare to straightforward and declarative statements. Automated fact-checker depends on natural language processing technology where it is more advanced in English than the other languages around the world. This is why a fact-checker is not widely used by the majority especially when it comes to recent, popular and unchecked claims or topics. Meanwhile, this research is believed to bring values to the society as this novel approach could help the users to detect fake news from recent and popular topics around the world by using its utility functions. This utility function has complied with the three criteria – trustable-source, data dispersion, and the similarity of queries to the title in checking a piece of news rather than the method used in an ordinary fact-checker.

3.3 Data Understanding

This phase begins with the data understanding before starting to create datasets and the data collecting process. A brief study across the internet had been done to indicate which platforms were more suitable in retrieving datasets for data mining purposes. There were many resources that were accessible from the internet such as Buzzmo, Alexa, Kaggle, and etc where these platforms provided a wide variety of data for learning purposes. Some interesting subsets can be perceived from data after a deep exploration process. This process is significant to ensure the data quality as the interesting components could help in forming the hypothesis in the research. Consequently, the interpretation of the Data Understanding phase will be illustrated further where a basic understanding from the available data was acquired.

3.3.1 Type of Fake News

“Fake news” are said to be a common yet familiar term where the current generation is constantly dealing with especially during the covid-19 outbreak. Due to the pandemic, consumption of information and news are increased exponentially and eventually caused many false information to go viral throughout the internet. It is crucial to identify the type of mis- and disinformation published on the online platforms as it helps the users to have better and clearer identification in detecting the fake news. There are seven types of fake news and each of them will be described in table 3.3.1.1 below.

Type of fake news	Description
1. Fabricated Content	Entirely untrue content
2. Manipulated Content	Contortion of actual information or media like photos, videos, etc. For example, sensationalized headlines that target on luring the readers' attention.
3. Imposter Content	Impersonate the authoritative parties. For example, using the trademarking or branding of an official organization or publisher.
4. Misleading Content	Deceiving use of information. For example, framing an issue or person which is totally not involved.

5. False context of connection	Genuinely accurate content is published with untruthful context. For example, the headline and the content of an article are not related.
6. Satire or Parody	Articles created by including current affairs or new items mixing with humorous features. For example, the use of irony, humour, sarcasm in an article to expose or criticize something.

Table 3.3.1.1 Type of fake news and respective description.

3.3.2 Data Creation and Collection

In this project, Google was selected as the best candidate for data collection as it provided wide, simple, and better search results regarding the queries that were inserted into the search function. This greatly increased the efficiencies in getting the recent popular topics, articles, news, and websites published all around the world that would be useful for the later phase. In this stage, three datasets were created called training, testing, and competitor's data. By understanding the differences of the type of fake news, the fabricated content type of fake news was selected as the standard for creating the fake query in a dataset. The models in this research will be dealing with fabricated content. Fabricated content is chosen as it could ease the detection of fake news within recently hot topics at the moment. Besides that, the real and fake queries were created according to current issues as well as recently popular topics which directly responded to the health domain news especially covid-19 news around the world. In addition, only the English version of queries were created as according to Johnson, J (2021), English was the most popular language being used over the worldwide internet users as of January 2020.

Thus, a list of real and fake queries following the standards mentioned above and comprised of fake and real labelling was manually created for training and testing the models. For training datasets, there are a total number of 60 queries related to current hot health topics with 30 real queries and 30 fake queries. The 10 samples of real and fake training data were show in figure 3.3.2.1.

No.	Query	Label
1	mco 1.0 start at 18 March 2020	REAL
2	mco 2.0 start at 13 jan 2021	REAL
3	cross-state is not allowed during mco	REAL
4	donald trump suggested that disinfectants, such as bleach, could be injected	REAL
5	dr noor hisham urge malaysian to wear mask	REAL
6	covid 19 vaccine is originated in Malaysia	FAKE
7	only male will be infected by covid 19	FAKE
8	Only Malaysians are immune to covid 19	FAKE
9	no people die from covid 19	FAKE
10	coronavirus will not cause death	FAKE

Figure 3.3.2.1 A sample of 10 rows of training queries regarding health domain issues with fake and real labelling.

As for testing data, it followed all the necessary requirements based on the training data. A total number of 30 testing data queries with 15 fake and real each were created for the testing and evaluating the accuracy of the model in modelling phase. 10 samples were shown in figure 3.3.2.2.

No.	Query	Label
1	most people will have side effects after getting covid-19 vaccination	REAL
2	stop panic buying during the pandemic says domestic trade minister	REAL
3	coronavirus vaccines' common side effects are headache, fatigue, fever etc.	REAL
4	covid 19 vaccines do not contain live coronavirus	REAL
5	Khairy Jamaluddin is the first patient who received Sinovac covid-19 vaccine in	REAL
6	hair straightener can kill the coronavirus	FAKE
7	covid19 vaccine will be jab at patient's head in order to achieve higher success	FAKE
8	eat rubbish waste can cure covid-19	FAKE
9	malaysia prime minister commit suicide after getting the covid19 vaccine	FAKE
10	prime minister commit that all covid19 vaccines will cause death	FAKE

Figure 3.3.2.2 A sample of 10 rows of testing queries regarding health domain issues with fake and real labelling.

A competitor dataset was collected for evaluating the models' capability on data with noise. This approach was targeting to detect fake news from recent and high demanding topics. Therefore, a competitor training dataset that was working in the same field – detecting fake news from GitHub was chosen for the model testing purpose. Hence, a total number of 60 rows of data with 30 real and fake news were extracted from the competitor's training data. A sample of 10 rows of competitor's data was present in figure 3.3.2.3. A competitor dataset was collected for evaluating the models' capability on data with noise. This approach was targeting to detect fake news from recent and

high demanding topics. Therefore, a competitor training dataset that was working in the same field – detecting fake news from GitHub was chosen for the model testing purpose. Hence, a total number of 60 rows of data with 30 real and fake news were extracted from the competitor’s training data. A sample of 10 rows of competitor’s data was present in figure 3.3.2.3.

id	title	author	text	label
0	House Dem Aide: We Didn't Even See Comey's Let...	Darrell Lucus	House Dem Aide: We Didn't Even See Comey's Let...	1
1	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life circles the rou...	0
2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, ...	1
3	15 Civilians Killed In Single US Airstrike Hav...	Jessica Purkiss	Videos 15 Civilians Killed In Single US Airstr...	1
4	Iranian woman jailed for fictional unpublished...	Howard Portnoy	Print lnAn Iranian woman has been sentenced to...	1
5	Jackie Mason: Hollywood Would Love Trump if He...	Daniel Nussbaum	In these trying times, Jackie Mason is the Voi...	0
6	Life: Life Of Luxury: Elton John's 6 Favorite ...	NaN	Ever wonder how Britain's most iconic pop plan...	1
7	Benoît Hamon Wins French Socialist Party's Pre...	Alissa J. Rubin	PARIS — France chose an idealistic, traditi...	0
8	Excerpts From a Draft Script for Donald Trump'...	NaN	Donald J. Trump is scheduled to make a highly ...	0
9	A Back-Channel Plan for Ukraine and Russia, Co...	Megan Twohey and Scott Shane	A week before Michael T. Flynn resigned as nat...	0

Figure 3.3.2.3 A sample of 10 rows of competitor’s data with different categories and ‘1’ (real) and ‘0’ (fake) labelling.

3.3.3 Credible Sources Collection

As stated earlier, there were three criteria fitting in this research approach in detecting fake news using data from Google search specifically on new and popular topics. One of the criteria used was checking the sources of results returned from Google search by inserting data queries created earlier as input queries. In consequence, a list of trustable website sources which are also known as credible publishers, parties as well as organizations was required for checking the truthfulness of every returned result from Google search by comparing the sources with the trustable sources list. Browsing through the authenticated news website had been done and most of the credible sources were collected into a file. Besides, checking on the web address of these “trustable” sites is significant in this step. Spelling errors in publisher companies' names or some strange domain extensions such as “.cool”, “.rich”, “.meme” may be giving a hint that the sources are suspicious and required to be verified before believing them. Some of the convincing sources that enable people to believe were official sites like World Health Organization or well-known publishers like New Straits Times etc. A total number of 132 trustable source links were collected at the moment. A sample of 10 credible sources collected was shown in figure 3.3.3.1.

No.	Website Links
1	https://kpksehatan.com/
2	https://www.cdc.gov/
3	https://www.infosihat.gov.my/
4	https://news.microsoft.com/
5	https://www.thestar.com.my/
6	https://www.freemalaysiatoday.com/
7	https://www.nst.com.my/
8	https://www.who.int/
9	https://edition.cnn.com/
10	https://www.scmp.com/

Figure 3.3.3.1 A sample of 10 rows of credible website links were collected across the internet.

3.4 Data Preparation

In this phase, the data are evaluated to ensure that they fit the scope of the project. Some cleaning techniques like filtering, merging, sorting and related tasks were carried out after the required datasets were fully prepared in the earlier phase. By giving examples, the regular expression such as paragraph break ‘\br’ or indentation in the raw data’s text are replaced with a blank space. In addition, the emoji, symbols, meaningless text, as well as special characters contained in the text, are also removed. On the other hand, other unnecessary labels such as id, author, text content, etc. as well as duplicated data are removed. All the datasets – training, testing, and competitor were standardized into a specified format with two categories which were Query and Label (‘FAKE’ and ‘REAL’). The cleaning task was essential as the datasets might contain useless variables or other components that will cause the model to be biased if they were included in the training and testing phases, eventually unclean data may affect the results of training and testing as well.

For example, a competitor’s data was required to transform into a specific format before it can be fit into the model. Labelling of ‘1’ and ‘0’ in the competitors’ dataset were changed to specified labels – ‘REAL’ and ‘FAKE’ where they indicated the categories of a piece of news. Besides that, only required data were remained for the model were the titles of articles and the genuineness of each of the article (‘Fake’ and ‘Real’ Labelling). The title of articles in the competitor’s data will be utilized in representing as queries or inputs that loaded into the Google search engine. Thus, the label of ‘title’ will be changed into ‘Query’. A final format of the competitor’s data was shown in figure 3.4.1.

No.	Query	Label
1	FLYNN: Hillary Clinton, Big Woman on Campus - Breitbart	FAKE
2	Why the Truth Might Get You Fired	REAL
3	15 Civilians Killed In Single US Airstrike Have Been Identified	REAL
4	Iranian woman jailed for fictional unpublished story about woman stoned t	REAL
5	Jackie Mason: Hollywood Would Love Trump if He Bombed North Korea ov	FAKE
6	A Back-Channel Plan for Ukraine and Russia, Courtesy of Trump Associates	FAKE
7	BBC Comedy Sketch "Real Housewives of ISIS" Causes Outrage	FAKE
8	US Officials See No Link Between Trump and Russia	REAL
9	In Major League Soccer, Argentines Find a Home and Success - The New Yo	FAKE
10	Wells Fargo Chief Abruptly Steps Down - The New York Times	FAKE

Figure 3.4.1 Screenshot of competitor's data after cleaning and transforming.

As for credible source lists, some minor data cleaning was done on them too. Removing the “/” sign symbol called slash and “https” characters were done on the trustable-source lists in order to have a systematized form when comparing the links returned from Google search. Aside from this, these links are neatly organized by removing unnecessary symbols, characters, and numbers. A final format of a credible source list was created and presented in the figure 3.4.2 below. As for credible source lists, some minor data cleaning was done on them too. Removing the “/” sign symbol called slash and “https” characters were done on the trustable-source lists in order to have a systematized form when comparing the links returned from Google search. Aside from this, these links are neatly organized by removing unnecessary symbols, characters, and numbers. A final format of a credible source list was created and presented in the figure 3.4.2 below.

No.	Website Links
1	kpkesehatan.com
2	www.cdc.gov
3	www.infosihat.gov.my
4	news.microsoft.com
5	www.thestar.com.my
6	www.freemalaysiatoday.com
7	www.nst.com.my
8	www.who.int
9	edition.cnn.com
10	www.scmp.com

Figure 3.4.2 Screenshot of trustable source list after cleaning.

3.5 Modelling

In this phase, the model implemented in this research will be explained in detail. There will be three models introduced in this section which are ID3 called Iterative Dichotomiser 3, Ensemble Learning, and Random Forests. The input that will be fed to the model will be pre-processed by a utility function that extract the 3 criteria from the raw inputs.

3.5.1 Operationalization of Criteria

Three criteria implied together with the utility function and models helped to indicated whether the approach objectives could be achieved. These three criteria were crucial in identifying the fake news using data collected from Google search specifically on trendy topics. Each of them played an important role as each of them represented a measure where they could bring meaningful information to this research. These criteria are stated and explained in the table 3.5.1.1 below.

Variable	Indicator
Trust	Criteria 1: Trustable Source Number of articles are from trustable source or site.
Spreading Rate	Criteria 2: Date Dispersion Average date difference between articles.
Accuracy of search result	Criteria 3: Similarity of Queries to Articles' Title Average similarity score.

Table 3.5.1.1 Description of criteria.

3.5.2 Utility Function

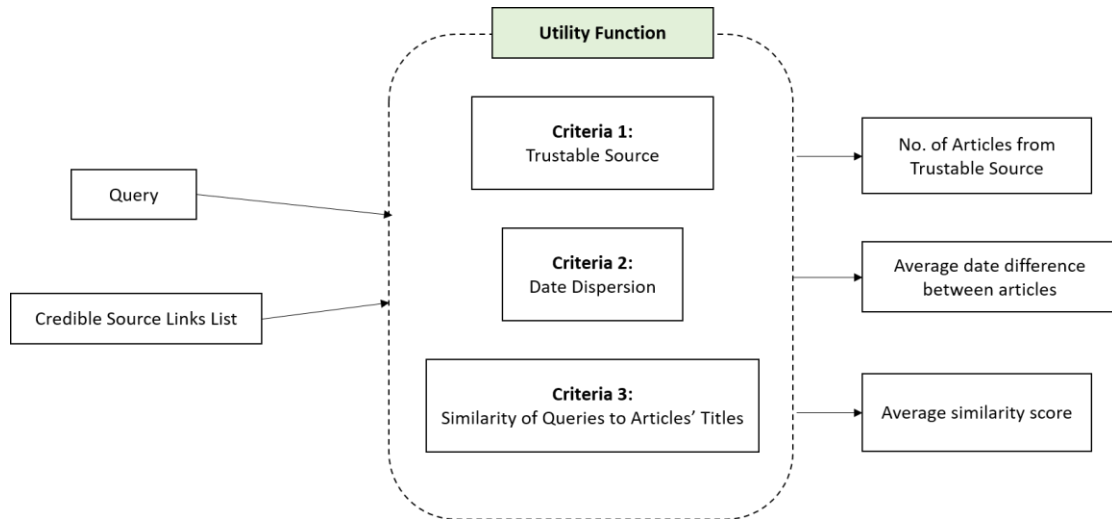


Figure 3.5.2.1 Utility Function.

Figure 3.5.2.1 provided a general overview of the utility function. At the early phase, three datasets – training, testing, and competitor’s data consisted of a number of queries will be inserted into this utility function. Top 10 Google search results for each query will be returned and collected, equally to say, one query will have 10 searched results. Each of these 10 results will then undergo the three criteria filtering function one by one in order to get the respective outputs. First of all, these 10 results will be filtered by checking the existence of the published date and title. For those results that do not fulfill these attributes, they will be removed, and the results that leftover will be passed to the next function. By going through the first criteria function, the remaining results’ website links will be used to compare with the credible source links list. This is to calculate the number of articles from these remaining results that were coming from the trustable website. This criterion was chosen because the query is more likely to be reliable if most of the websites retrieved from the query were from credible sources. Next, the same remaining results were passed to the second criteria function - date dispersion for calculating the average date difference between articles. This is to evaluate the spreading rate of an issue as the probability of fake news occurring will be higher when the spreading rate is high. The dates were sorted in descending order where the first date will be used to subtract the second date, the second date will be used to

subtract the third date, and so on. The date difference of each subtraction will be added and divided by the number of remaining results for getting the average date difference of the query. The last criterion was then performed using results that were from trustable sources. Only trustable websites' titles will be compared with the query to calculate the average similarity scores between these attributes. The number of similar words between query and title will be calculated first before divided by the total length of the title. If there are multiple instances of the same word, only one will be counted when counting the interception. This process will be repeated for all the trustable websites. A sum will be obtained, and the average will be calculated. This criterion was used for indicating the accuracy of search results corresponding to the query. If the similarity score is low, the query is more likely to be fake. In consequence, the number of articles from trustable sources, an average of date dispersion between articles, the average similarity score of queries to titles were obtained from the utility function.

3.5.3 ID3, Iteration Dichotomiser 3 Model

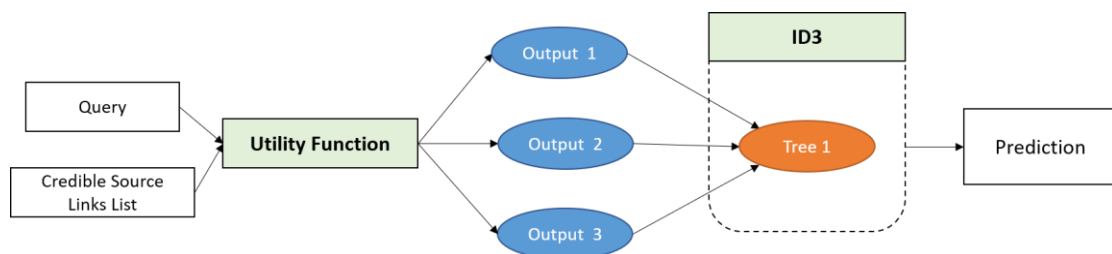


Figure 3.5.3.1 Architecture diagram for ID3.

The fundamental block for three of the models is the tree decision classifier. The ID3 is constructed with just a block of the tree decision classifier. The outputs from the utility function that will be pumped into the ID3 refer to the number of articles from trustable sources, the average date difference between articles and the average of similarity score of queries to titles. The outputs will be served as the features for the classifier to predict the label of the query.

3.5.4 Ensemble Learning Model

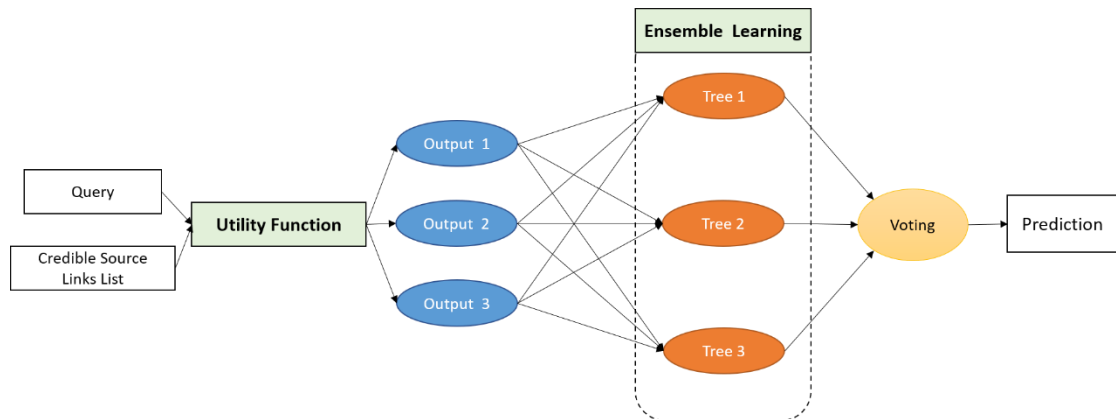


Figure 3.5.4.1 Architecture diagram for Ensemble Learning.

The second model is constructed with 3 tree decision classifiers. During the training phase, the classifiers are trained with 1/3 of the training data each. Since three of the classifiers are receiving different data during the training phase, the decisions made by the classifiers on the same input would be different. Therefore, a voting system is required to combine three of the outputs from the classifiers. Majority rules method is implemented in the voting system.

3.5.5 Random Forest Model

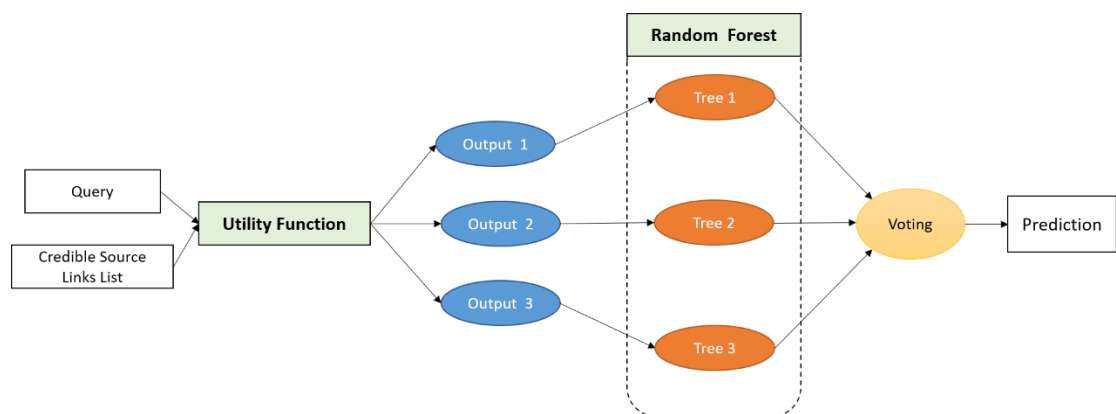


Figure 3.5.5.1 Architecture diagram for Random Forest.

In the third model, each of the classifiers will be responsible for one of the criteria. A voting system is also required in this method. The voting method will follow the majority rules where the majority decisions will be taken as the output of the model.

3.6 Evaluation

At this point, an evaluation on the models implemented will be carried out. This is to determine the quality and effectiveness of the designed approach as well as the final outputs. A benchmark will be created by generating the label 'FAKE' and 'REAL' randomly for testing the testing data. The accuracy of the benchmark will be used to determine the effectiveness of the models. Once the models have a poorer accuracy than the benchmark, fine tuning on the models is required. For example, changing the calculation methods for the utility function.

On the other hand, competitor's data will be used to test the effectiveness of the model on the real-world data which are from different domain topics. As the domain of data is not just limited to the health domain, the accuracy for competitor's data might vary.

3.7 Deployment

The results evaluated based on the models will be determined and the project's objectives are expected to be achieved in detecting the fake news specifically on recent and trendy topics. The deployment of this research will be uploading the datasets created and the utility function as well as the three models' source code into the code hosting platform like GitHub or Kaggle repository.

3.8 Tools and Technologies Used

Python is chosen as the programming language for the proposed model in this research. A couple of libraries were used in this approach as these libraries provided a lot of efficient tools for machine learning. In particular, googlesearch library is crucial to this research because google search results are needed for modelling and evaluating. Other libraries such as nltk (Natural Language Toolkit) for human language data processing, htmldate for getting article's timestamp, url_metadata, numpy, pandas, datetime, sklearn, bs4 (BeautifulSoup) for pulling data from HTML and XML files, random etc. in this project. All the experiments conducted in research will be implemented in

Google Collab, a collaboratory that allows users to code and execute python code through the browser.

Python Libraries Version	
Beautiful Soup, bs4	0.0.1
Datetime	3.7.10
Googlesearch	2.0.3
Natural Language Toolkit, nltk	3.2.5
Numpy	1.19.5
Pandas	1.1.5
Sklearn	0.0
Statistics	1.0.3.5
Url_metadata	0.1.6

Table 3.8.1 Python libraries version.

Google Collab Hardware Specs: CPU-only VMs	
Specifications	Descriptions
CPU	2.30GHz Intel(R) Xeon(R)
No. CPU Cores	2
RAM	12GB
Disk Space	25GB

Table 3.8.2 Google Collab Specifications.

3.9 Implementation Issues and Challenges

Some issues had come across in this project. In this research, data creation and collection are the most important yet time-consuming tasks. This is said so because the datasets created and collected may affect the change of result during evaluation indirectly, so it can be described as the core of this research. At the time, this task is time-consuming as the training and testing datasets were self-created. This is because the fabricated content type of fake queries is a criterion in creating the datasets. Due to

the limitation of the approach at the moment, only fabricated content type of fake queries is considered in determining the effectiveness of the three criteria used.

Besides that, only popular and current issues in health domain-related queries were accepted before inserting into the utility function. Therefore, some changing and updating the queries were done as they required to fit this domain, else they would affect the accuracy of the models. Despite knowing the difference between the fake news, it is still very hard to create a false statement that met that specific requirement. Thus, rephrasing or replacing of queries are done multiple times to ensure that all the data are fabricated content.

Besides that, the calculations or similarity functions used have been changed multiple times to improve the accuracy of the models. However, in order to know the performance of the functions, testing had to be done. This process is repetitive and inevitable. Moreover, the calculations had to be desk checked to ensure that the output received meet the expectation and the calculation is done as intended.

Meanwhile, when collecting the credible source links, a simple verification will be done on each of the website addresses to ensure they are from reliable publishers. Aside from this, there are numerous amounts of websites around the world, however, due to the time constraint and limited manpower, validating all those websites is impossible. Therefore, the website list created in this research might be more biased towards websites from Malaysia. Besides that, the utility function required humans to supervise it during the training phase as it will be stuck when retrieving data from the website. Hence, manual interrupting is required.

3.10 Timeline

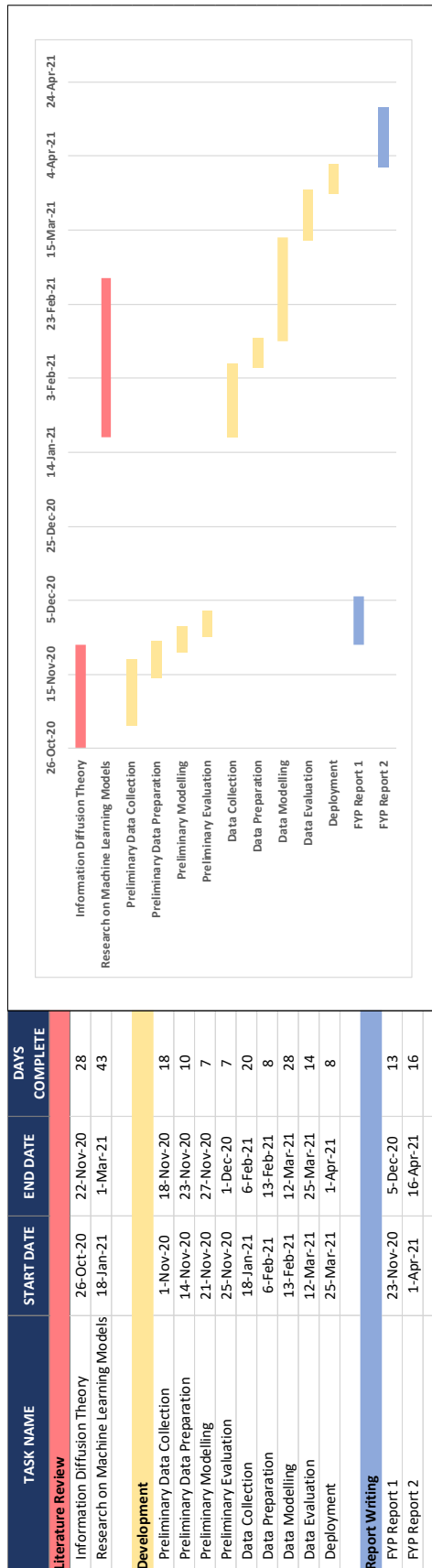


Figure 3.10 Gantt chart.

CHAPTER 4: EXPERIMENTAL RESULTS

4.1 Sample Data

4.1.1 Training Data

A total of 60 rows of training data were created manually. As mentioned in the section above, 60 queries comprised of 30 real and fake news, specifically on recent and popular health domain news were created as training data. The majority of the queries were related to the current outbreak topics – covid-19 news and all the data were in the English language. Thus, constantly surfing the net is required to identify the trendy and hot news that were published recently. Figure 4.1.1.1 below shows the final version of 15 samples from 60 rows of data with the label after undergoing necessary refinement.

No.	Query	Label
1	mco 1.0 start at 18 March 2020	REAL
2	mco 2.0 start at 13 jan 2021	REAL
3	cross-state is not allowed during mco	REAL
4	donald trump suggested that disinfectants, such as bleach, could be	REAL
5	dr noor hisham urge malaysian to wear mask	REAL
6	Dr Adham Baba advice Malaysians to drink warm water to kill coro	REAL
7	Kelantan MP suggested the use of antidotes based upon Rukyah ve	REAL
8	Minor side effects after having any vaccine are common	REAL
9	Pfizer-biontech is one of the covid19 vaccine	REAL
10	older people is free from covid 19 virus	FAKE
11	covid 19 vaccine is originated in Malaysia	FAKE
12	only male will be infected by covid 19	FAKE
13	Only Malaysians are immune to covid 19	FAKE
14	no people die from covid 19	FAKE
15	coronavirus will not cause death	FAKE

Figure 4.1.1.1 Sample of training dataset.

4.1.2 Testing Data

For testing data, a total number of 30 rows of queries where 15 queries were fake and real each were created for testing the model. These 30 queries fulfill the requirements stated for training data. For instance, the current issues or topics like covid-19 vaccines invention, safeness of covid-19 vaccine, side-effects, and many coronavirus-health related concerns were used as the queries. A sample testing data was shown in the figure 4.1.2.1 below.

No.	Query	Label
1	most people will have side effects after getting covid-19 vaccination	REAL
2	stop panic buying during the pandemic says domestic trade minister	REAL
3	coronavirus vaccines' common side effects are headache, fatigue, f	REAL
4	covid 19 vaccines do not contain live coronavirus	REAL
5	Khairy Jamaluddin is the first patient who received Sinovac covid-19	REAL
6	Sinovac covid-19 vaccine is made in China	REAL
7	Pfizer-BioNTech vaccine is originated from US	REAL
8	only female will die after receiving covid19 vaccine	FAKE
9	covid19 patients will turn into zombies	FAKE
10	people who violet standard operation procedure will not be fined	FAKE
11	hair straightener can kill the coronavirus	FAKE
12	covid19 vaccine will be jab at patient's head in order to achieve high	FAKE
13	eat rubbish waste can cure covid-19	FAKE
14	malaysia prime minister commit suicide after getting the covid19 va	FAKE
15	prime minister commit that all covid19 vaccines will cause death	FAKE

Figure 4.1.2.1 Sample of testing dataset.

4.1.3 Competitors Data

While for competitor's data, it was downloaded from the GitHub repository as the field of this chosen dataset was related to fake news detection using machine learning too. However, the domain of this dataset was not just limited to recent health issues, it comprised of many other news categories like politics, sport, business, economics, etc. Hence, this selected dataset will be used for testing the model in order to obtain a significant evaluation if the model is testing an outsource data. Since the data content was different, essential data like the title (articles' titles) and label (truthfulness of article) in the datasets were extracted and transformed into the format that we wanted. The title of the article will be treated as the queries and the label will be remained. Besides that, data cleaning was done on these extracted data stated in section 3.4. 60 rows of data were extracted, and a sample of this competitor's dataset was shown in figure 4.1.3.1 after transforming and cleaning process.

No.	Query	Label
1	FLYNN: Hillary Clinton, Big Woman on Campus - Breitbart	FAKE
2	Why the Truth Might Get You Fired	REAL
3	15 Civilians Killed In Single US Airstrike Have Been Identified	REAL
4	Iranian woman jailed for fictional unpublished story about woman stoned t	REAL
5	Jackie Mason: Hollywood Would Love Trump if He Bombed North Korea o	FAKE
6	A Back-Channel Plan for Ukraine and Russia, Courtesy of Trump Associates	FAKE
7	BBC Comedy Sketch "Real Housewives of ISIS" Causes Outrage	FAKE
8	US Officials See No Link Between Trump and Russia	REAL
9	In Major League Soccer, Argentines Find a Home and Success - The New Yo	FAKE
10	Wells Fargo Chief Abruptly Steps Down - The New York Times	FAKE

Figure 4.1.3.1 Sample of competitor's dataset.

4.1.4 Trustable Website sources

A total number of 132 trustable website sources were prepared in this research. This website list will be utilized in calculating the number of articles returned were from credible origins when the queries searched on Google. Some familiar and convincing sources that included in this website list were the world health organization official website, America famous news publishers – CNN, The New York Times, USA Today, etc, Malaysia popular news publishers – News Strait Times, The Star, Malaysiakini, etc. and many others authentic sources around the world. A final version of this website list was shown below after cleaning the unnecessary symbols, characters, and extensions.

No.	Website Links
1	kpkesehatan.com
2	www.infosihat.gov.my
3	news.microsoft.com
4	www.thestar.com.my
5	www.freemalaysiatoday.com
6	www.nst.com.my
7	www.who.int
8	edition.cnn.com
9	thestar.com.my
10	thesundaily.my
11	www.dailymail.co.uk
12	www.bbc.com
13	www.nytimes.com
14	www.usatoday.com
15	www.nydailynews.com

Figure 4.1.4.1 Sample of trustable website sources.

4.2 Utility Function and Models Implementation

Before training three of the models, a benchmark was created by using a random generator to create the labels on testing data. The accuracy of random generator labels on testing data was used as the benchmark to evaluate all the models. A benchmark value of 0.5 was obtained with this method. If the models' accuracy is more than or at least equal to this benchmark value, this could indicate that the models' ability to classify the news are questionable. After getting the benchmarks, the model training phase was initiated by using the training dataset on the utility function. All the model's implementations will be performed on Google Colab using Python with version Py 3.6.7.

First of all, the training data with queries and labels shown in figure 4.1.1.1 will be loaded into the utility function. By using the queries, data from google search results could be acquired. Each query will return at most 10 results from the Google search engine every iteration. From the data retrieved, necessary information like website links, titles, metadata, timestamp, etc. could be pulled out using python libraries. The filtering process will be executed on these data to ensure the existence of website links, timestamps and titles were significant for future modelling. The remaining data must have these three attributes before reaching the three criteria functions. Equally saying that not every query will have 10 results leftovers after the filtering process. Then, the three criteria functions will be performed on these remaining data to get the desired outputs. The first criterion function – trustable source was called to calculate the number of data was from credible sources by comparing with the trustable websites list created earlier. Secondly, the date dispersion function was then called for calculating the average of date distance between the data remained. The date of data that remained were sort in descending order where the first date was the largest while the last date was the smallest one among them. Subtractions were performed between the first and second date, continue for second and third, third and fourth, and so on. All the date differences calculated will then be summed up and divided to get an accuracy value. The dispersion function calculation can be described as:

$$\text{Average} = \frac{(\text{first} - \text{second}) + (\text{second} - \text{third}) + (\text{third} - \text{fourth}) + \dots}{\text{number of webstie which have title and date}}$$

Last but not least, the last criterion function which was the similarity of queries to titles was then called for calculating the average of similarity score. The number of data were from a trustable source calculated at the first criterion represented an attribute in this function. Only data that come from trustable sources were used to calculate the similarity of queries to title as this function was to indicate the accuracy of google search results using our queries. Therefore, the similarity score formula can be described as following:

$$\text{Similarity Score} = \frac{\text{number of words matched between query and title}}{\text{total number of words in a title}}$$

After getting all the similarity scores, an average of these scores will be computed. Lastly, a table of data was obtained after the utility function was successfully executed.

Three outputs of each query which were the number of articles from trustable source, an average of date difference between articles, and average similarity score of queries to titles were acquired and display in the figure 4.2.1 below.

Label	Query	No of Trustable Source	No of finalURLlen	Date Dispersion	Query Vs Title
REAL	Foreigners living in Malaysia will receive free	3	8	0.125	0.167200855
FAKE	prime minister of malaysia is confirmed as c	4	7	1.714285714	0.276515152
REAL	Prime minister Muhyiddin Yassin became the	4	8	2.625	0.4625
REAL	Malaysia approves Sinovac, AstraZeneca cov	3	8	2.75	0.781144781
REAL	mco 2.0 start at 13 jan 2021	2	10	3.1	0.168016194
FAKE	covid 19 vaccine is originated in Malaysia	2	7	3.714285714	0.087121212
FAKE	100k covid-19 confirmed cases on 2 feb mal	2	8	6.125	0.266666667
REAL	cross-state is not allowed during mco	5	10	6.3	0.121051693
FAKE	dr noor hisham died from covid-19	10	10	7.2	0.207842501
REAL	donald trump suggested that disinfectants, s	7	10	8.1	0.257096965
REAL	halal vaccine is recognized in Malaysia	4	8	9.75	0.170192308
FAKE	people who received any covid19 vaccine wi	5	10	10.3	0.109225742
REAL	Pfizer-biontech is one of the covid19 vaccine	4	10	10.3	0.250315657
REAL	health workers at high risk of exposure and c	4	7	13.42857143	0.216666667
FAKE	children are immune to covid 19 virus	0	9	14.88888889	0
REAL	Dr Adham Baba advice Malaysians to drink w	4	10	30.1	0.171428571
FAKE	coronavirus won't spread	3	10	31	0.132154882
REAL	spm of 2020 is postponed to 22 feb 2021	3	10	33	0.363596491
FAKE	Thermal scanners can detect covid 19	2	10	33.3	0.183333333
REAL	dr noor hisham urge malaysian to wear masl	7	8	34.625	0.120185078
FAKE	coronavirus spread through social media	1	9	35.55555556	0.25
REAL	stpm of 2020 is postponed to 3 marc 2021	7	10	36.6	0.201560449
FAKE	coronavirus will not cause death	2	8	39	0

Figure 4.2.1 Table of data after executing Utility Function.

Based on the figure above, the data that will be given to three of the models for training are label, the number of trustable websites (No. of Trustable Source), date dispersion between articles (Date Dispersion), and the similarity score of the query to titles (Query Vs Title). For the first model – ID3, it will receive all 60 of the training data during the training phase. On the other hand, each of tree in the Ensemble Learning model will receive one-third of the training data. While for the Random Forest model, only one of the outputs from three of the criteria will be received by each tree.

After the models were trained, a number of 30 rows of testing data will be loaded into utility function for information extraction similar to what had been done to the training data. A table with the same format as the output of training data will be obtained after finish executing the utility function with testing data. With the information obtained, the three models can be used to predict the query accordingly. For testing data, the “label” attribute is not needed for the models as the three models will predict the label for testing data. Therefore, only the number of trustable websites (No. of Trustable Source), date dispersion between articles (Date Dispersion), and the similarity score of queries to titles (Query Vs Title) are needed for the models. In the ID3 model, all data comprised with the three criteria will be inputted for predicting the label of 30 rows of

testing data. On the other hand, for the second model – ensemble learning, 30 rows of testing data comprised with the three criteria will be inputted into each of the trees since this model have three-segmented trees. A voting method by following the majority will be implemented to these trees for determining the final prediction outputs. There will be 30 predicted labels as the outputs of this model. At the same time, the third model which is the random forest model was executed too. Since the trees in this model are trained with one criterion each, data is loaded into their respective tree. Similarly, a voting method is used for indicating the final prediction for all data. At the end of the process, a total of 30 predictions will be obtained from this model also.

After receiving all the predictions from each of the models, the predictions will be compared to the actual label of the testing data. Accuracy, f1 score, precision, and recall will be calculated for performance evaluation of the models. The models' performance should be higher than the benchmark else the model will be retrained with a different seed. When the model is ready to receive real-world input, the competitor's data will be inputted to the utility function and predictions similar to the testing phase will be carried out by three of the models. 60 rows that are sampled from the competitor data will be used in this phase to stimulate the performance of the model in the deployment stage. The accuracy, f1 score, precision, and recall will also be calculated for model evaluation.

4.2.1 Performance Evaluation

Testing Data	Accuracy	F1 Score	Precision	Recall	Confusion Matrix	
ID3 Model	0.70	0.69	0.71	0.67	TN: 11	FP: 4
					FN: 5	TP: 10
Ensemble Learning Model	0.73	0.71	0.77	0.67	12	3
					5	10
Random Forest Model	0.5	0.44	0.50	0.40	9	6
					9	6

Table 4.2.1.1 Performance evaluation of testing data.

Based on the performance evaluation obtained, accuracy, f1 score, precision, recall, and confusion matrix were obtained among the models. ID3 model has an accuracy of 0.70,

while ensemble learning has an accuracy of 0.73 whereas the last model – random forest model has an accuracy of 0.5. As we can see, ensemble learning is the best model that gave the highest accuracy value of 0.73 among the three models in predicting the label of a query correctly. Anyhow the rest models are still considerable as their accuracies are at least above than or equal to the benchmark with a value of 0.5.

Competitor's Data	Accuracy	F1 Score	Precision	Recall	Confusion Matrix	
ID3 Model	0.33	0.46	0.39	0.57	TN: 3	FP: 27
					FN: 13	TP: 17
Ensemble Learning Model	0.32	0.39	0.35	0.43	6	24
					17	13
Random Forest Model	0.30	0.36	0.33	0.40	6	24
					18	12

Table 4.2.1.2 Performance evaluation of competitor's data.

Since the three models were ready to accept real-world inputs, the performance evaluation base on the table above was obtained. By referring to the table above, we can see that the accuracy among the three models is not convincing as all of them are lower than the benchmark value which is 0.5. Accuracy around 0.30 from each of the models was acquired and reasons for these low accuracies are needed to identify. Some reasons and justification for explaining the low accuracies were gathered. First, the news category of competitor's data is one of the factors that affected the model accuracy. This is said because the competitor's data contained a lot of other categories of news such as business, politics, entertainment, etc which are not specified in health-related topics. Besides that, majority of the competitor's data are in old and untrendy topics compared to our training data which from recent and popular covid-19 health topics. Due to these reasons, the accuracies among the models were expected to be low since the domain of competitors and training data are different. The output is expected or predictable before performing this experiment. This experiment clearly showcases the ability of the models are limited to the domain of the training data.

CHAPTER 5: CONCLUSION

5.1 Project Review

This research combined three criteria with the utility function and worked together with ID3, Ensemble Learning, and Random Forest modality to build a new approach that can detect fake news using data from Google search specifically on recent and popular topics. The three criteria which are the number of articles from trustable source, data dispersion between articles, and similarity of queries to titles applied to the utility function play significant roles in improving the accuracy for letting the utility function extracts these important yet meaningful data for training. Other than this, the inclusion of several modalities enabled the proposed model to achieve remarkable performance while only using a small amount and self-created information. This research combined three criteria with the utility function and worked together with ID3, Ensemble Learning, and Random Forest modality to build a new approach that can detect fake news using data from Google search specifically on recent and popular topics. The three criteria which are the number of articles from trustable source, data dispersion between articles, and similarity of queries to titles applied to the utility function play significant roles in improving the accuracy for letting the utility function extracts these important yet meaningful data for training. Other than this, the inclusion of several modalities enabled the proposed model to achieve remarkable performance while only using a small amount and self-created information.

Every single phase of the methodologies in this research had been completed and the performance evaluation had given us a consequential view on this novel approach. Since the testing accuracy was above the benchmark and also higher than the competitor's accuracy, hereby the motive and the objectives are said to be achieved. This research had brought contribution towards the society in identifying a piece of fake news, especially from recent trendy news by just using three simple criteria. More than that, new and recent fake news could be detected easily without checking them at a traditional and ordinary fact-checker. Every single phase of the methodologies in this research had been completed and the performance evaluation had given us a consequential view on this novel approach. Since the testing accuracy was above the benchmark and also higher than the competitor's accuracy, hereby the motive and the objectives are said to be achieved. This research had brought contribution towards the

society in identifying a piece of fake news, especially from recent trendy news by just using three simple criteria. More than that, new and recent fake news could be detected easily without checking them at a traditional and ordinary fact-checker.

In conclusion, the objectives of this research were achieved with a 70% of accuracy rate by the inclusion of various criteria with a self-created utility function and implemented them into few models. This novel approach is able to estimate the authenticity of the news from data using Google search specifically on recent and trendy topics. In conclusion, the objectives of this research were achieved with a 70% of accuracy rate by the inclusion of various criteria with a self-created utility function and implemented them into few models. This novel approach is able to estimate the authenticity of the news from data using Google search specifically on recent and trendy topics.

5.2 Future Work

Subsequent tasks could be done to improve the overall performance in this research. Due to the time limit, the experimental on more criteria rather than using three criteria in the utility function could not be carried out in this project. Hence, in the future, more criteria will be applied to the utility function to increase the ability on detecting fake news. Besides that, the number of datasets should be increased in the future so that the models can be trained and tested on a larger amount of data. This could help to create a more robust model. Since this approach only achieved approximately 70% accuracy at the moment, it is believed to improve in the future by implementing other models that fit into this research. There might be a more suitable model than the models used in this project that could classify the news better. By doing so, the evaluation on the model performance is believed to be improved and the accuracy is believed to be higher than 70%.

Bibliography

- About Facebook. (2019). *Company Info | About Facebook*. [online] Available at: <https://about.fb.com/company-info/> [Accessed 10 Sep. 2020].
- Ali, J., Khan, R., Ahmad, N. and Maqsood, I. (2012). Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)*, 9(5), p.272.
- Allcott, H. and Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, 31(2), pp.211–236.
- Allcott, H., Gentzkow, M. and Yu, C. (2018). Trends in the diffusion of misinformation on social media. *Research & Politics*, 6(2), p.205316801984855.
- Azlan, A.A., Hamzah, M.R., Sern, T.J., Ayub, S.H. and Mohamad, E. (2020). Public knowledge, attitudes and practices towards COVID-19: A cross-sectional study in Malaysia. *PLOS ONE*, 15(5), p.e0233668.
- Beebe, M. (2018). *Research Guides: Fake News, Misinformation & Disinformation: Types of Misinformation & Disinformation*. [online] shawneesu.libguides.com. Available at: <https://shawneesu.libguides.com/c.php?g=651556&p=4570051>.
- Bruns, A. and Highfield, T. (2013). POLITICAL NETWORKS ON TWITTER. *Information, Communication & Society*, 16(5), pp.667–691.
- Bruns, A. and Highfield, T. (n.d.). *Political Networks on Twitter: Tweeting the Queensland State Election*. [online] Available at: <http://snurb.info/files/2013/Political%20Networks%20on%20Twitter.pdf>.
- Budak, C., Agrawal, D. and El Abbadi, A. (2011). Structural trend analysis for online social networks. *Proceedings of the VLDB Endowment*, 4(10), pp.646–656.
- Cataldi, M., Caro, L. and Schifanella, C. (2010). *Emerging Topic Detection on Twitter based on Temporal and Social Terms Evaluation*. [online] Available at: <http://www.ai.univ-paris8.fr/~cataldi/papers/mdm-kdd2010.pdf> [Accessed 11 Sep. 2020].

BIBLIOGRAPHY

- Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C., Brugnoli, E., Schmidt, A., Zola, P., Zollo, F. and Scala, A. (2020). *The COVID-19 Social Media Infodemic*. [online] Available at: <https://arxiv.org/pdf/2003.05004.pdf>.
- Data Science Project Management. (n.d.). *CRISP-DM*. [online] Available at: <https://www.datascience-pm.com/crisp-dm-2/>.
- De Vries, E.L.E. (2019). When more likes is not better: the consequences of high and low likes-to-followers ratios for perceived account credibility and social media marketing effectiveness. *Marketing Letters*, 30(3–4), pp.275–291.
- Dietterich T.G. (2000) Ensemble Methods in Machine Learning. In: Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science, vol 1857. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-45014-9_1
- Gomez-Rodriguez, M., Balduzzi, D., De, B., Schölkopf, B. and De, B. (2011). *Uncovering the Temporal Dynamics of Diffusion Networks*. [online] pp.561–568. Available at: <http://snap.stanford.edu/class/cs224w-readings/rodriguez11diffusion.pdf> [Accessed 10 Sep. 2020].
- Gomez-Rodriguez, M., Leskovec, J. and Krause, A. (2012). Inferring Networks of Diffusion and Influence. *ACM Transactions on Knowledge Discovery from Data*, 5(4), pp.1–37.
- Gomez-Rodriguez, M., Leskovec, J. and Schölkopf, B. (2013). *Structure and Dynamics of Information Pathways in Online Media*. [online] Available at: <https://cs.stanford.edu/~jure/pubs/infopath-wsdm13.pdf> [Accessed 10 Sep. 2020].
- Guille, A., Hacid, H., Favre, C. and Zighed, D.A. (2013). Information diffusion in online social networks. *ACM SIGMOD Record*, 42(2), pp.17–28.
- Hua, J. and Shaw, R. (2020). Corona Virus (COVID-19) “Infodemic” and Emerging Issues through a Data Lens: The Case of China. *International Journal of Environmental Research and Public Health*, 17(7), p.2309.

BIBLIOGRAPHY

- J, M. (2020). *Malaysia: social media penetration 2020*. [online] Statista. Available at: <https://www.statista.com/statistics/883712/malaysia-social-media-penetration/#:~:text=As%20of%20January%202020%2C%20about>.
- Johnson, J. (2021). *Internet: most common languages online 2019* / Statista. [online] Statista. Available at: <https://www.statista.com/statistics/262946/share-of-the-most-common-languages-on-the-internet/>.
- Kouzy, R., Abi Jaoude, J., Kraitem, A., El Alam, M.B., Karam, B., Adib, E., Zarka, J., Traboulsi, C., Akl, E. and Baddour, K. (2020). Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter. *Cureus*.
- Lu, R. and Yang, Q. (2012). Trend Analysis of News Topics on Twitter. *International Journal of Machine Learning and Computing*, 2(3), pp.327–332.
- Nadia, M.B. and Daniel, L.S. (n.d.). (PDF) *Aging in an Era of Fake News*. [online] ResearchGate. Available at: https://www.researchgate.net/publication/341496718_Aging_in_an_Era_of_Fake_News.
- Navlani, A. (2018). *Decision Tree Classification in Python*. [online] Available at: <https://www.datacamp.com/community/tutorials/decision-tree-classification-python>.
- Page, L., Brin, S., Motwani, R. and Winograd, T. (1999). The PageRank Citation Ranking: Bringing Order to the Web. - Stanford InfoLab Publication Server. *Stanford.edu*. [online] Available at: <http://ilpubs.stanford.edu:8090/422/> [Accessed 10 Sep. 2020].
- Peng, W., Chen, J. and Zhou, H. (n.d.). *An Implementation of ID3 ---Decision Tree Learning Algorithm*. [online] . Available at: <https://rhuang.cis.k.hosei.ac.jp/Miccl/AI-2/DecisionTree2.pdf> [Accessed 14 Apr. 2021].
- Ramayah, U. (2020). *COVID-19: Public awareness on the dangers of spreading fake news increases*. [online] Astroawani.com. Available at: <https://www.astroawani.com/berita-malaysia/covid19-public-awareness-on->

BIBLIOGRAPHY

- the-dangers-of-spreading-fake-news-increases-245663 [Accessed 10 Sep. 2020].
- Sakkaf, Y. (2020). *Decision Trees for Classification: ID3 Algorithm Explained*. [online] Medium. Available at: <https://towardsdatascience.com/decision-trees-for-classification-id3-algorithm-explained-89df76e72df1>.
- Simpson, E. and Conner, A. (2020). *Fighting Coronavirus Misinformation and Disinformation*. [online] Center for American Progress. Available at: <https://www.americanprogress.org/issues/technology-policy/reports/2020/08/18/488714/fighting-coronavirus-misinformation-disinformation/>.
- Stieglitz, S., Dang-Xuan, L., Bruns, A. and Neuberger, C. (2014). Social Media Analytics. *Business & Information Systems Engineering*, [online] 6(2), pp.89–96. Available at: <https://link.springer.com/article/10.1007/s12599-014-0315-7> [Accessed 10 September 2019].
- Wang, M., Zuo, W. and Wang, Y. (2015). *A Multilayer Naïve Bayes Model for Analyzing User's Retweeting Sentiment Tendency*. [online] Computational Intelligence and Neuroscience. Available at: <https://www.hindawi.com/journals/cin/2015/510281/#EEq9> [Accessed 10 Sep. 2020].
- Wang, S., Liu, H., Shu, K., Mahudeswaran, D. and Lee, D. (2018). *FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media Fake News Detection and Mitigation on Social Media View project Feature engineering for outlier detection View project Deepak Mahudeswaran FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media*.
- Welle (www.dw.com), D. (2020). *Spread of coronavirus fake news causes hundreds of deaths | DW | 11.08.2020*. [online] DW.COM. Available at: <https://www.dw.com/en/coronavirus-misinformation/a-54529310> [Accessed 10 Sep. 2020].

BIBLIOGRAPHY

- Wirth, R. and Hipp, J. (2000). *CRISP-DM: Towards a Standard Process Model for Data Mining*. [online] Available at: <http://www.cs.unibo.it/~montesi/CBD/Beatriz/10.1.1.198.5133.pdf> [Accessed 11 Sep. 2020].
- www.who.int. (2020). *Munich Security Conference*. [online] Available at: <https://www.who.int/dg/speeches/detail/munich-security-conference> [Accessed 10 Sep. 2020].
- Ying, L. (2019). *10 Twitter Statistics Every Marketer Should Know in 2020 [Infographic]*. [online] my.oberlo.com. Available at: <https://my.oberlo.com/blog/twitter-statistics#:~:text=Here> [Accessed 10 Sep. 2020].
- Yusof, A.N.M., Muuti, M.Z., Ariffin, L.A. and Tan, M.K.M. (2020). Sharing Information on COVID-19: the ethical challenges in the Malaysian setting. *Asian Bioethics Review*.

APPENDIX A: Poster

A NOVEL APPROACH TO DETECT FAKE NEWS USING DATA FROM GOOGLE SEARCH SPECIFICALLY ON RECENT AND POPULAR TOPICS

PANG HUEY JING
Bachelor of Computer Science (Hons)

Dr Ooi Boon Yaik
Project Supervisor
Faculty of Information and Communication Technology



Universiti Tunku Abdul Rahman

ABSTRACT

The Internet has become an important source of health information to the public worldwide during the covid-19 pandemic. An enormous amount of fake or misleading health information has been widely spread across social media platforms during the pandemic. Hence, a novel approach is used to detect fake news using data scraped from Google search specifically on recent and popular topics. By utilizing the criteria which are checking the source, date dispersion of articles and the accuracy of search result, this research model can act as a current issue related news checker that allows the public to filter out fake news published across the internet. A self-created utility function is created for extracting the features of data from Google search with the three criteria stated. A 70% of accuracy rate as well as the objectives were achieved using different models in this approach.



PROJECT OBJECTIVES



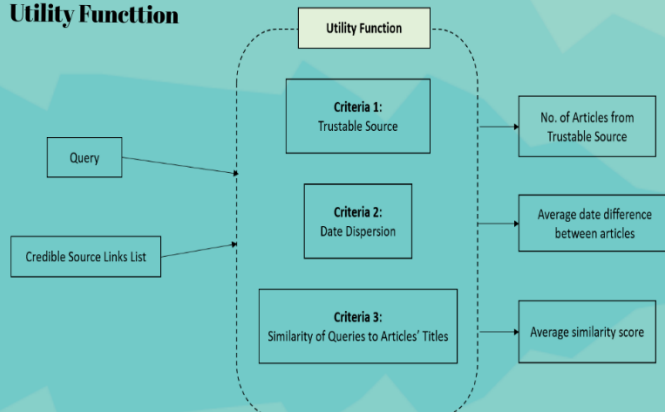
To detect fake news by implementing the three criteria.



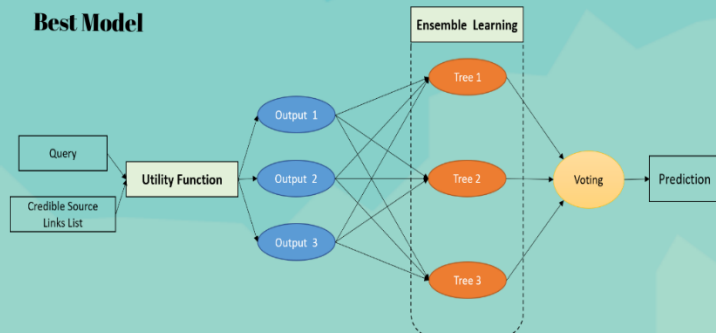
To determine the effectiveness of the three criteria in this research.

METHODOLOGY

Utility Function



Best Model



Model Performance	Testing Data	Accuracy	F1 Score	Precision	Recall	Confusion Matrix	
	ID3 Model	0.70	0.69	0.71	0.67	TN: 11	FP: 4
Ensemble Learning Model	0.73	0.71	0.77	0.67	12	3	
					5	10	
Random Forest Model	0.5	0.44	0.50	0.40	9	6	
					9	6	

APPENDIX B: Final year project weekly report

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: Trimester 3, Year 3	Study week no: 2
Student Name & ID: Pang Huey Jing 17ACB02304	
Supervisor: Ts Dr Ooi Boon Yaik	
Project Title: A Novel Approach To Detect Fake News Using Data From Google Search Specifically On Recent And Popular Topics	

1. WORK DONE

[Please write the details of the work done in the last fortnight.]

- i. Figured out methods to collect data from Google.
- ii. Implement required tasks using Google Collab.
- iii. Performed data scraping to collect top 10 results returned from Google.

2. WORK TO BE DONE

- i. Figure out possible criteria that help to detect possible fake news.
- ii. Investigate the metadata collected from Google.

3. PROBLEMS ENCOUNTERED


- i. Methodologies and concepts have to change due to change in supervisor.

4. SELF EVALUATION OF THE PROGRESS

- i. Self- assigned tasks are able to complete within the expected time.



 Supervisor's signature



 Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: Trimester 3, Year 3	Study week no: 4
Student Name & ID: Pang Huey Jing 17ACB02304	
Supervisor: Ts Dr Ooi Boon Yaik	
Project Title: A Novel Approach To Detect Fake News Using Data From Google Search Specifically On Recent And Popular Topics	

1. WORK DONE

[Please write the details of the work done in the last fortnight.]

- i. Created trustable websites list.
- ii. Created trustable websites compare function to validate the source of data.
- iii. Created dataDispersion function to calculate the date distance of the data.

2. WORK TO BE DONE

- i. Creating a utility function to calculate the accuracy of training data with criteria.
- ii. Investigating the usefulness of trustable websites criteria.
- iii. Investigating the usefulness of date dispersion criteria.

3. PROBLEMS ENCOUNTERED

- i. Accuracy score of the criteria is lower than expectations.

4. SELF EVALUATION OF THE PROGRESS

- iv. Self- assigned tasks are able to complete within the expected time.



Supervisor's signature



Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: Trimester 3, Year 3	Study week no: 6
Student Name & ID: Pang Huey Jing 17ACB02304	
Supervisor: Ts Dr Ooi Boon Yaik	
Project Title: A Novel Approach To Detect Fake News Using Data From Google Search Specifically On Recent And Popular Topics	

1. WORK DONE

[Please write the details of the work done in the last fortnight.]

- i. Compare the source of training data with trustable website lists.
- ii. Created a utility function that calculate the accuracy of training data.
- iii. Tried WordNet with NLTK method.

2. WORK TO BE DONE

- i. Change training dataset's queries.
- ii. Increase training datasets.
- iii. Change the date dispersion calculation methods.
- iv. Add new criteria which is comparing the data's queries and titles.

3. PROBLEMS ENCOUNTERED

- i. Accuracy of each criteria is lower than expectations.
- ii. WordNet with NLTK is unable to implement in this project as there are some libraries error occurred.

4. SELF EVALUATION OF THE PROGRESS

- iv. Self- assigned tasks are able to complete within the expected time.



Supervisor's signature



Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: Trimester 3, Year 3	Study week no: 8
Student Name & ID: Pang Huey Jing 17ACB02304	
Supervisor: Ts Dr Ooi Boon Yaik	
Project Title: A Novel Approach To Detect Fake News Using Data From Google Search Specifically On Recent And Popular Topics	

1. WORK DONE

[Please write the details of the work done in the last fortnight.]

- I. Calculated the accuracy of the three criteria.
- II. Interpreted the output of these three datasets.
- III. Investigate the effectiveness of the criteria for filtering out possible fake news.

2. WORK TO BE DONE

- I. Increase the number of data for training and competitor's data.
- II. Create testing data for testing.

3. PROBLEMS ENCOUNTERED

- I. Quality of data are needed to improve.
- II. Difficulties in differentiate the type of fake news.

4. SELF EVALUATION OF THE PROGRESS

- I. Self- assigned tasks are able to complete within the expected time.



Supervisor's signature



Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: Trimester 3, Year 3	Study week no: 10
Student Name & ID: Pang Huey Jing 17ACB02304	
Supervisor: Ts Dr Ooi Boon Yaik	
Project Title: A Novel Approach To Detect Fake News Using Data From Google Search Specifically On Recent And Popular Topics	

1. WORK DONE

[Please write the details of the work done in the last fortnight.]

- IV. Calculated the accuracy of the three criteria.
- V. Compared the accuracy of training, testing and competitor's data.
- VI. Interpreted the output of these three datasets.

2. WORK TO BE DONE

- I. Change the domain of the testing data
- II. Train the decision tree classifier.
- III. Compare the accuracy of the classifier on both testing and competitor's data.

3. PROBLEMS ENCOUNTERED

- I. Accuracy of the classifier was undesirable because of some minor error occurred in labelling.
- II. Domain of the datasets do interfere the output of classifier.

4. SELF EVALUATION OF THE PROGRESS

- I. Self- assigned tasks are able to complete within the expected time.



Supervisor's signature



Student's signature

FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

Trimester, Year: Trimester 3, Year 3	Study week no: 12
Student Name & ID: Pang Huey Jing 17ACB02304	
Supervisor: Ts Dr Ooi Boon Yaik	
Project Title: A Novel Approach To Detect Fake News Using Data From Google Search Specifically On Recent And Popular Topics	

1. WORK DONE

[Please write the details of the work done in the last fortnight.]

- I. Train the decision tree classifier.
- II. Evaluate the classifier
- III. Compare the accuracy of the classifier on both testing and competitor's data.

2. WORK TO BE DONE

- I. Prepare Final Year Project 2 Report
- II. Update contents for every chapters of the report

3. PROBLEMS ENCOUNTERED

- I. Required time for reviewing more literatures.

4. SELF EVALUATION OF THE PROGRESS

- I. Self- assigned tasks are able to complete within the expected time.



Supervisor's signature



Student's signature

APPENDIX C: Plagiarism Check Result

The screenshot shows the Turnitin Feedback Studio interface. The document title is "Pang Huey Jing | A NOVEL APPROACH TO DETECT FAKE NEWS USING DATA FROM GOOGLE SEARCH SPECIFICALLY ON RECENT AND POPULAR TOPICS". The document content includes:

CHAPTER 1 INTRODUCTION

CHAPTER 1: INTRODUCTION

1.1 Problem Statement and Motivation

The coronavirus disease pandemic has caused a huge burden to all the countries around the world and resulted more than millions of deaths. It is believed that false information are existed and are being spread uninhibitedly over the social media platforms at a noticeable speed when the public health or healthcare officials rushed to identify the virus that may spread. Misinformation can propagate across the internet without the need for any professional verification, constraints as well as any scientific proofs. (Kouzy et al., 2020). Hence, the spreading of false information in a country over social media or website could happen anytime, anywhere especially when there is a lack of information or evidence associated with a topic.

Coronavirus is a new virus and had brought chaos to the global in a blink of an eye. Therefore, there has been limited data published in the world regarding population

On the right side, the "Match Overview" panel shows a total similarity score of 8%. The matches are listed as follows:

Match Number	Source	Similarity
1	Submitted to South Ba... Student Paper	1%
2	Adrien Guille, Hakim H... Publication	1%
3	Ceren Budak, Divyakant... Publication	1%
4	link.springer.com Internet Source	1%
5	downloads.hindawi.com Internet Source	<1%
6	Hunt Allcott, Matthew ... Publication	<1%
7	Manuel Gomez-Rodrig... Publication	<1%
8	Submitted to Hacettep... Student Paper	<1%
9	docplayer.net Internet Source	<1%

Page: 1 of 48 | Word Count: 13243 | Text-only Report | High Resolution On

The screenshot shows the Turnitin Originality Report for the document "A NOVEL APPROACH TO DETECT FAKE NEWS USING DA... By Pang Huey Jing". The report was processed on 16-Apr-2021 10:11 +08. The similarity index is 8%.

Similarity Index: 8%

Similarity by Source:

Source Type	Percentage
Internet Sources	4%
Publications	5%
Student Papers	3%

The report lists the following matches:

- 1% match (student papers from 01-Dec-2011): Submitted to South Bank University on 2011-12-01
- 1% match (publications): Adrien Guille, Hakim Hacid, Cecile Favre, Djamel A. Zighed. "Information diffusion in online social networks". ACM SIGMOD Record, 2013
- 1% match (publications): Ceren Budak, Divyakant Agrawal, Amr El Abbadi. "Structural trend analysis for online social networks". Proceedings of the VLDB Endowment, 2011
- 1% match (Internet from 22-Aug-2019): https://link.springer.com/article/10.1007%2Fs12599-014-0315-7
- <1% match (): http://downloads.hindawi.com
- <1% match (publications): Hunt Allcott, Matthew Gentzkow, Chuan Yu. "Trends in the diffusion of misinformation on social media". Research & Politics, 2019
- <1% match (publications): Manuel Gomez-Rodriguez, Jure Leskovec, Andreas Krause. "Inferring Networks of Diffusion and Influence". ACM Transactions on Knowledge Discovery from Data, 2012
- <1% match (Internet from 09-Jun-2017): http://docplayer.net
- <1% match (Internet from 23-Nov-2020): https://towardsdatascience.com/machine-learning-basics-decision-tree-from-scratch-part-i-4251bfa1b45c?gi=52365820af40
- <1% match (student papers from 24-Apr-2018)

Universiti Tunku Abdul Rahman			
Form Title: Supervisor's Comments on Originality Report Generated by Turnitin for Submission of Final Year Project Report (for Undergraduate Programmes)			
Form Number: FM-IAD-005	Rev No.: 0	Effective Date: 01/10/2013	Page No.:



**FACULTY OF INFORMATION AND
COMMUNICATION TECHNOLOGY**

Full Name(s) of Candidate(s)	Pang Huey Jing
ID Number(s)	17ACB02304
Programme / Course	Bachelor Of Computer Science (Honours)
Title of Final Year Project	A Novel Approach To Detect Fake News Using Data From Google Search Specifically On Recent And Popular Topics

Similarity	Supervisor's Comments (Compulsory if parameters of originality exceeds the limits approved by UTAR)
Overall similarity index: <u> 8 </u> % Similarity by source Internet Sources: <u> 4 </u> % Publications: <u> 5 </u> % Student Papers: <u> 3 </u> %	
Number of individual sources listed of more than 3% similarity: <u> - </u>	
Parameters of originality required, and limits approved by UTAR are as Follows: (i) Overall similarity index is 20% and below, and (ii) Matching of individual sources listed must be less than 3% each, and (iii) Matching texts in continuous block must not exceed 8 words <i>Note: Parameters (i) – (ii) shall exclude quotes, bibliography and text matches which are less than</i>	

Note Supervisor/Candidate(s) is/are required to provide softcopy of full set of the originality report to Faculty/Institute

Based on the above results, I hereby declare that I am satisfied with the originality of the Final Year Project Report submitted by my student(s) as named above.

Alex

Signature of Supervisor

Signature of Co-Supervisor

Name: Ts Dr Ooi Boon Yaik

Name: _____

Date: 16/04/2021

Date: _____



UNIVERSITI TUNKU ABDUL RAHMAN

FACULTY OF INFORMATION & COMMUNICATION TECHNOLOGY (KAMPAR CAMPUS)

CHECKLIST FOR FYP2 THESIS SUBMISSION

Student Id	17ACB02304
Student Name	PANG HUEY JING
Supervisor Name	TS DR OOI BOON YAIK

TICK (√)	DOCUMENT ITEMS
	Your report must include all the items below. Put a tick on the left column after you have checked your report with respect to the corresponding item.
√	Front Cover
√	Signed Report Status Declaration Form
√	Title Page
√	Signed form of the Declaration of Originality
√	Acknowledgement
√	Abstract
√	Table of Contents
√	List of Figures (if applicable)
√	List of Tables (if applicable)
√	List of Symbols (if applicable)
√	List of Abbreviations (if applicable)
√	Chapters / Content
√	Bibliography (or References)
√	All references in bibliography are cited in the thesis, especially in the chapter of literature review
√	Appendices (if applicable)
√	Poster
√	Signed Turnitin Report (Plagiarism Check Result - Form Number: FM-IAD-005)

*Include this form (checklist) in the thesis (Bind together as the last page)

<p>I, the author, have checked and confirmed all the items listed in the table are included in my report.</p> <p style="text-align: center;"><i>Ar</i></p> <hr style="width: 20%; margin: 0 auto;"/> <p>(Signature of Student) Date: 16/04/2021</p>	<p>Supervisor verification. Report with incorrect format can get 5 mark (1 grade) reduction.</p> <p style="text-align: center;"><i>Alex</i></p> <hr style="width: 20%; margin: 0 auto;"/> <p>(Signature of Supervisor) Date: 16/04/2021</p>
---	---