

COVID-19 FAKE NEWS DETECTION MODEL ON
SOCIAL MEDIA DATA USING MACHINE LEARNING
TECHNIQUES

KELVIN LIEW KAI XUAN

BACHELOR OF COMPUTER SCIENCE
(GRAPHIC & MULTIMEDIA TECHNOLOGY) WITH
HONOURS

UNIVERSITI MALAYSIA PAHANG

UNIVERSITI MALAYSIA PAHANG

DECLARATION OF THESIS AND COPYRIGHT

Author's Full Name : KELVIN LIEW KAI XUAN

Date of Birth

Title : COVID-19 FAKE NEWS DETECTION MODEL ON SOCIAL MEDIA DATA USING MACHINIE LEARNING TECHNIQUES

Academic Session : 2022/2023

I declare that this thesis is classified as:

- CONFIDENTIAL (Contains confidential information under the Official Secret Act 1997)*
- RESTRICTED (Contains restricted information as specified by the organization where research was done)*
- OPEN ACCESS I agree that my thesis to be published as online open access (Full Text)

I acknowledge that Universiti Malaysia Pahang reserves the following rights:

1. The Thesis is the Property of Universiti Malaysia Pahang
2. The Library of Universiti Malaysia Pahang has the right to make copies of the thesis for the purpose of research only.
3. The Library has the right to make copies of the thesis for academic exchange.

Certified by:

(Student's Signature)

(Supervisor's Signature)

Nur Shazwani Binti Kamarudin

New IC/Passport Number
Date: 19 January 2023

Name of Supervisor
Date: 19 January 2023

NOTE : * If the thesis is CONFIDENTIAL or RESTRICTED, please attach a thesis declaration letter.

THESIS DECLARATION LETTER (OPTIONAL)

Librarian,
Perpustakaan Universiti Malaysia Pahang,
Universiti Malaysia Pahang,
Lebuhraya Tun Razak,
26300, Gambang, Kuantan.

Dear Sir,

Author's Name

Thesis Title

Reasons (i)

(ii)

(iii)

CLASSIFICATION OF THESIS AS RESTRICTED

Please be informed that the following thesis is classified as RESTRICTED for a period of three (3) years from the date of this letter. The reasons for this classification are as listed below.

Thank you.

Yours faithfully,



(Supervisor's Signature)

Date: 19 January 2023

Stamp:

DR. NUR SHAZWANI KAMARUDIN
PENSYARAH KANAN
UNIVERSITI MALAYSIA PAHANG
26600 PEKAN, PAHANG.
TEL : 09-424 4736

Note: This letter should be written by the supervisor, addressed to the Librarian, *Perpustakaan Universiti Malaysia Pahang* with its copy attached to the thesis.



SUPERVISOR'S DECLARATION

I hereby declare that I have checked this thesis and in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Bachelor of Computer Science in Graphics & Multimedia Technology.

A handwritten signature in black ink, appearing to be 'Nur Shazwani', written over a horizontal line.

(Supervisor's Signature)

Full Name : Nur Shazwani binti Kamarudin

Position : Senior Lecturer

Date : 19 January 2023

(Co-supervisor's Signature)

Full Name :

Position :

Date :



STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Universiti Malaysia Pahang or any other institutions.

A handwritten signature in black ink, appearing to read 'Kelvin', is written above a horizontal line.

(Student's Signature)

Full Name : KELVIN LIEW KAI XUAN

ID Number : CD19040

Date : 19 January 2023

COVID-19 FAKE NEWS DETECTION MODEL ON SOCIAL MEDIA DATA
USING MACHINIE LEARNING TECHNIQUES

KELVIN LIEW KAI XUAN

Thesis submitted in fulfillment of the requirements
for the award of the degree of
Bachelor of Computer Science in Graphic and Multimedia Technology

Faculty of Computing

UNIVERSITI MALAYSIA PAHANG

JANUARY 2023

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to Dr Nur Shazwani binti Kamarudin, my final project supervisor at the Faculty of Computer Science, Universiti Malaysia Pahang, for allowing me to work under her supervision and for providing useful guidance throughout the final project process. Her patience, tolerance, enthusiasm, sincerity and motivation helped me to complete my final project successfully. She guided me on how to do the project and how to communicate the results as matter-of-factly as possible. It has been a great honour and privilege to work and study under her guidance. I am grateful for all she has done for me. I am grateful for her tolerance and compassion during our discussions about the project work and the preparation of my thesis.

In addition, I like to thank my parents for their love, dedication, concern and education and efforts for my future. I truly appreciate their support, compassion, dedication and perseverance, without which I would not be the person I am today. And to my friends for their unwavering support in moving forward and cheering me on when I feel sad.

Finally, I would like to express my gratitude to all those who have helped me with my research, whether directly or indirectly.

ABSTRAK

Laman media sosial seperti Instagram, Twitter dan Facebook telah menjadi bahagian yang sangat diperlukan dalam rutin harian. Laman media sosial ini adalah instrumen yang kuat untuk menyebarkan berita, gambar, dan jenis maklumat lain. Namun, sejak munculnya pandemi COVID-19 pada bulan Disember 2019, banyak artikel dan tajuk utama mengenai wabak COVID-19 telah muncul di media sosial. Media social sering digunakan untuk menyebarkan bahan atau maklumat palsu. Maklumat yang tidak betul ini boleh mengelirukan pengguna, mungkin menimbulkan kebimbangan. Sukar untuk mengatasi penyebaran maklumat yang meluas. Hasilnya, sangat penting untuk mengembangkan model untuk mengenali berita palsu dalam aliran berita. Set data, yang akan menjadi sintesis berita berkaitan COVID-19 dari banyak media sosial dan sumber berita, digunakan untuk pengkategorian dalam karya ini. Penanda diambil dari data teks yang tidak tersusun yang dikumpulkan dari pelbagai sumber. Kemudian, untuk menghilangkan beban komputasi menganalisis semua ciri dalam set data, pemilihan ciri dilakukan. Akhirnya, untuk mengkategorikan set data yang berkaitan dengan covid -19, algoritma pembelajaran mesin canggih dilatih. Support Vector Machine (SVM), Naïve Bayes (NB), dan Decision Tree (DT) adalah model pembelajaran mesin yang dipersembahkan. Akhirnya, banyak langkah digunakan untuk menilai algoritma ini.

ABSTRACT

Social media sites like Instagram, Twitter and Facebook have become an indispensable part of the daily routine. These social media sites are powerful instruments for spreading news, photographs, and other sorts of information. However, since the emergence of the COVID-19 pandemic in December 2019, many articles and headlines concerning the COVID-19 epidemic have surfaced on social media. Social media is frequently used to disseminate fraudulent material or information. This disinformation may confuse consumers, perhaps causing worry. It is hard to counter the widespread dissemination of disinformation. As a result, it is critical to develop a model for recognising fakes news in the news stream. The dataset, which would be a synthesis of COVID-19-related news from numerous social media and news sources, is utilised for categorization in this work. Markers are retrieved from unstructured textual data gathered from a variety of sources. Then, to eliminate the computational burden of analysing all of the features in the dataset, feature selection is done. Finally, to categorise the covid -19 related dataset, multiple cutting-edge machine learning algorithms were trained. Support Vector Machine (SVM), Naïve Bayes (NB), and Decision Tree (DT) are the machine learning models presented. Finally, numerous measures are used to evaluate these algorithms.

TABLE OF CONTENT

DECLARATION	
TITLE PAGE	
ACKNOWLEDGEMENTS	ii
ABSTRAK	iii
ABSTRACT	iv
TABLE OF CONTENT	v
LIST OF TABLES	viii
LIST OF FIGURES	ix
CHAPTER 1 INTRODUCTION	1
1.1 INTRODUCTION	1
1.2 PROBLEM STATEMENT	3
1.3 OBJECTIVE	4
1.4 SCOPE	4
1.5 SIGNIFICANCE PROJECT	5
1.6 REPORT ORGANIZATION	5
CHAPTER 2 LITERATURE REVIEW	6
2.1 INTRODUCTION	6
2.2 PREVIOUS RESEARCH WORKS	6
2.3 SUMMARY	9

CHAPTER 3 METHODOLOGY	10
3.1 INTRODUCTION	10
3.2 RESEARCH FRAMEWORK	10
3.2.1 Data Collection	11
3.2.2 Data Pre-processing	11
3.2.3 Data Classification	11
3.2.4 Experimental Result	12
3.2.5 Evaluating Model	12
3.3 PROJECT REQUIREMENT	13
3.3.1 Input	13
3.3.2 Output	13
3.3.3 Process Description	14
3.3.4 Constraints And Limitations	15
3.4 PROPOSED DESIGN	16
3.4.1 Flowchart	16
3.4.2 Flow Explanation	17
3.5 DATA DESIGN	18
3.5.1 Data Description	18
3.6 PROOF OF INITIAL CONCEPT	19
3.6.1 Machine Learning Algorithms	19
3.7 HARDWARE AND SOFTWARE EQUIPMENT	22
3.8 POTENTIAL USE OF PROPOSED SOLUTION	23
CHAPTER 4 IMPLEMENTATION, RESULTS AND DISCUSSION	24
4.1 INTRODUCTION	24
4.2 RESULT	24

4.3	DICUSSION	35
CHAPTER 5 CONCLUSION		36
5.1	OBJECTIVE REVISITED	36
5.2	LIMITATION	36
5.3	FUTURE WORK	37
REFERENCE		38
APPENDIX A		40

LIST OF TABLES

Table 2.1	Comparison Between Existing Research Papers	8
Table 3.1	Hardware equipment	22
Table 3.2	Software equipment	22
Table 4.1	Accuracy, Precision, Recall and F1_Score of Machine Learning Algorithms	25

LIST OF FIGURES

Figure 3.1	Framework of research	10
Figure 3.2	Accuracy, Precision, Recall, F1 formula	13
Figure 3.3	Flowchart of research	16
Figure 3.4	Dataset Description	18
Figure 3.5	Table content of dataset	18
Figure 3.6	Support Vector Machin	19
Figure 3.7	Naïve Bayes	20
Figure 3.8	Decision Tree	21
Figure 4.1	Comparison of Machine Learning Algorithms on Accuracy, Precision, Recall, F1_Score	26
Figure 4.2	Result of Accuracy for Different Algorithms in both Training and Testing	27
Figure 4.3	Result of Precision for Different Algorithms in both Training and Testing	28
Figure 4.4	Result of Recall for Different Algorithms in both Training and Testing	29
Figure 4.5	Result of F1_score for Different Algorithms in both Training and Testing	30
Figure 4.6	Fake and True News common words	31
Figure 4.7	Positive words which is true news for the text	31
Figure 4.8	Negative words which is fake news for the text	32
Figure 4.9	Confusion matrix of Decision Tree in Training	33
Figure 4.10	Confusion matrix of Support Vector Machine in Testing	34

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

Social media platforms like Facebook, Twitter, Instagram, and others have risen in the twenty-first century, allowing information to travel swiftly. Users on social media can publish whatever they wish, regardless of the provenance and credibility of the published material, posing problems to information dependability assurance. Each social media user could have as many accounts as they wish. With the COVID-19 epidemic, millions of posts or news are being sent out every day, with some detrimental repercussions for people and society. For example, the propagation of false information concerning COVID-19 patient data or symptoms that have yet to be confirmed. People might quickly become panicked as a result of social media's fake news and disinformation. The phrases fake news and disinformation are closely related and are sometimes used interchangeably. An automated false news detection system is required, which will rely on human annotation, machine learning, and natural language processing (Oshikawa et al., 2020).

Since its emergence in December 2019, there's been several articles and media articles on the COVID-19 outbreak in internet, traditional print, and digital media. These sources provide data from both credible and untrustworthy clinical sources. Furthermore, news from these outlets spreads swiftly. Spreading a false information might prompt nervousness, undesirable openness to clinical cures, stunts for computerized showcasing, and may prompt destructive elements. As a result, a model for detecting fake news in the news stream is critical. Fake news identification is a new technique, and there is still more

work to be done to attain specific levels of accuracy, particularly in news and information management (Oshikawa et al., 2020).

In this day and age, machine learning (ML) algorithms and natural language processing (NLP) algorithms are critical tools for identifying fake news. For several advanced work in computing science, NLP has grown into a fundamental pillar for comprehending and applying the ideas of text summarization, text categorization, sentiment analysis, and opinion mining (Agrawal et al., 2021). While machine learning (ML) relies on algorithms and data to construct models with minimum human participation (Alenezi & Alqenaie, 2021). In this study, the fake news detection model will be built using Python programming. The Natural Language Toolkit (NLTK) will be used to aggregate all of the social media data imputed by NLP and ML. It is difficult to process the large variety of data sources as there are many words in a social media post or news and it's not possible to read all the contents and pick out covid-10-related contents one by one. Therefore, this study proposes to use sentiment methods based on social media datasets to detect COVID-19 news to achieve sufficient accuracy, which requires the use of a keyword approach to capture COVID-19 news to determine which text messages are meaningful and to detect those that are fakes.

1.2 PROBLEM STATEMENT

Consequently, the country has not only been susceptible due to the spread of the COVID -19 outbreak, but has also exacerbated the situation as the 'epidemic' of false news has grown increasingly difficult to manage, causing instability within the population. This behaviour, whether for 'fun' or for evil, plainly exposes community to terror, as well as causing panic and insecurity.

Fake news concerning COVID-19 appeared to be quickly circulating on social media. Similar trends have been observed during previous epidemics, such as with the recent Ebola, yellow fever, and Zika outbreaks. It is a concerning trend since even a single encounter with disinformation might raise the accuracy of its perceptions. As a result, there is a strong need for an exploratory investigation of the social media platform account responsible for the COVID-19 disinformation, the distribution of COVID-19 misinformation on social media, and the substance of the COVID-19 misinformation spreading on social media platform.

The issue fake news detection on COVID-19 is harnessing auxiliary information on social media is a difficulty in in of itself, because users' social interactions with fake news generate vast amounts of imprecise, unorganized, and inconsistent data (Shu et al., 2017). Because these auxiliary messages are mixed with some real messages and some fake ones, it is difficult for users to distinguish between COVID-19 fake news. Information about COVID-19 fake news is also growing on social media over time, and all these unreliable information are slowly causing people to panic. To detect COVID-19 fake news more successfully, it is critical to depend on machine learning (ML) and natural language processing (NLP) to build a fake news detection model..

1.3 OBJECTIVE

- i. To study the use of social media data for COVID-19 fake news identification.
- ii. To design a detection model use machine learning and natural language processing on COVID-19 fake news.
- iii. To develop and evaluate a system or methodology that can utilise data from previous news stories to forecast if a news is phoney or not.

1.4 SCOPE

User scope

- i. Users a minimum of 13 years and a maximum of 90 years
- ii. Users who stay in Malaysia

System scope

- i. The dataset for this study comes from the social media platform Twitter, Facebook, WhatsApp etc.

Development scope

- i. Using content like text from social media data, Python and Machine Learning will be utilised to build a detection model to detect fake news on COVID-19.
- ii. Using dataset is based on COVID-19 fake news only.

1.5 SIGNIFICANCE PROJECT

Local Resident

- Local residents are better able to avoid being misinformed or fake news about COVID-19 from social media.

Outsider

- Those who are migrant workers or tourists will not be deterred from coming to Malaysia by COVID-19 fake news.
- Those who are migrant workers or tourists will not be influenced by COVID-19 fake news to not come to Malaysia.

1.6 REPORT ORGANIZATION

Chapter 1 describes on the introduction of the project including the background, problem statement, objectives, scope and significance of the project.

Chapter 2 is the literature review of the project that discuss on reviewing three existing systems, technologies and software that related to the project.

Chapter discusses on the methodology that will be implemented in the project which are the research framework discusses the process of the research and the project requirement discusses the scope of the research.

Chapter 4 is the discusses the result and the result obtained from the implemented method and technique of this research

Chapter 5 discusses the project's conclusion, project constraints, and future work for this project system.

CHAPTER 2

LITERATURE REVIEW

2.1 INTRODUCTION

Fake news refers to the purposeful dissemination of disinformation via traditional media or social media. Misinformation spreads very quickly. When one fake news website is taken down, another soon replaces it. Furthermore, because fake news travels so rapidly, it can be difficult to discern from factual reporting. Anyone can download articles from websites, exchange information, then re-share it with others, and by the end of the day, the incorrect information has migrated so far away from its originating location that it is unrecognisable. Disinformation has spread so far from its initial source that it can no longer be distinguished from genuine news. There have been several research conducted on the detection of fake news on social media. This study of the literature will compare previous research on fake news.

2.2 PREVIOUS RESEARCH WORKS

The machine learning ensemble strategy is used in this work (Villagrancia Octaviano, 2021) to solve the fake news detection challenge. Language query and word count (LIWC) feature sets are used in conjunction with ensemble approaches. ISOT Fake News Dataset and Kaggle are two datasets taken from the World Wide Web. Linguistic features are used to transform particular textual attributes into digital inputs for training models. The LIWC tool extracts features from the corpus, and the tool extract 93 distinct characteristics from each given text. Because all characteristics retrieved using the tool are numerical, categorical variables do not need to be encoded. Certain variables (such as percentages) have a 0 to 100 range while others have an unrestricted range. Scaling

range for values of various features must be ensure in the range of (0,1). Each dataset was partitioned into 70/30 training and test sets.

Another study (Jain et al., 2019), the authors used SVM pre-existing best-matched technique with Naïve Bayes since SVM is best suited for binary classification. The suggested approach includes three primary modules: news aggregator, news authenticator, and news recommendations system. News aggregator sites may see news and information from many sources, gather information, classify it in tags/categories, and organise it suitably for simpler usage. Then, the utilized news authenticator assists authors in determining if the given news is fake or real. Because of the multidimensional character of fake news, the recommended technique for this study is a combination of Naïve Bayes classifier, Support Vector Machines, and semantic investigation. The variables in Naïve Bayes classifier independent of each other to generate the model and machine learning algorithm is to supervise using Bayes' theorem. Classification was conducted by maximum a posteriori is derived. Furthermore, SVM extracts the binary class from the dataset and classifies the data into two groups, true or false, to determine the hyperplane which best splits the dataset into two classes. This study's accuracy by utilizing a mix approach or method is 93.50%, which is better than using only one method in prior research to determine the accuracy of fake news.

The third study (Monti et al., 2019) proposed of Twitter social network of fake news detection using geometric deep learning approach. This study's proposed model was trained supervised on a stream of tagged fake and authentic news that flowed on Twitter between 2013 and 2018. First, forecast false or true-class probabilities using a four-layer graph Convolutional Neural Network (CNN) with two convolution layers (each layer has a 64-dimensional output feature graph) and two connected layers. The study looked at two alternative settings for false news detection: Url-wise and Cascade-wise, and both used the same architecture. They attempted to predict the true or false tag of a URL that containing a new article using all of the Twitter cascades it created, and each URL generated an average of 141 crosstabs. In the second setting, they assume a single URL cascade and attempt to forecast the tags attached with that URL. The dataset was divided into training or validation which 80% of URLs and test sets which are 20% of URLs for URL-wise and cascade-wise configuration.

Table 2.1: Comparison Between Existing Research Papers

Elements	Research 1	Research 2	Research 3
Research and Author	Fake News Detection Using Machine Learning Ensemble Methods (Villagracia Octaviano, 2021)	A Smart System For Fake News Detection Using Machine Learning (Jain et al., 2019)	Fake News Detection on Social Media using Geometric Deep Learning (Monti et al., 2019)
Domain	Machine learning ensemble technique used to solve the challenge of detecting false news.	Using machine learning classifiers on Social Media data for fake news detection to stop the spread of activities such as mob lynching	Geometric deep learning technique for detecting fake news on the Twitter social network
Technique	<ul style="list-style-type: none"> - Linguistic Inquiry and Word Count (LIWC) - Three learning models of voting classifier <ol style="list-style-type: none"> 1. Logistic regression, random forest, and k-nearest neighbors (KNN) 2. Linear SVM, classification and regression trees (CART) 	Naïve Bayes classifier, Support Vector Machines (SVM), Natural Language Processing (NLP) and semantic investigation	Geometric deep learning - In particular, computer vision, speech analysis, and natural language processing, where geometric deep learning has had a significant impact.
Data	Two types of datasets extracted from World Wide Web <ol style="list-style-type: none"> 1. ISOT Fake News Dataset 2. Kaggle Dataset 	<ul style="list-style-type: none"> - From social media like Facebook, Instagram, Twitter, WhatsApp - Most popular sites that give semi-structured news data are Google News, Feedly, and News360. 	Twitter URLs – separate into two classes, training set and test set.
Advantages	To achieve maximum accuracy, learning models were trained and parameterized.	Naïve Bayes - By merely counting the class distribution, the Naïve Bayes assumption dramatically decreases the computing cost. SVM - Generally, it is quite exact and works exceedingly well on semi-structured datasets.	<ul style="list-style-type: none"> - Accomplishes very high precision and resilient behaviour in a variety of tough circumstances requiring vast amounts of large-scale real data. - Learn task-specific attributes from data automatically.

Disadvantages	KNN - efficiency or speed of algorithm declines very fast when dataset grows CART - a small change in the dataset might cause the tree structure to become unstable, causing variation.	- Naïve Bayes classifier estimations can be wrong in some cases - The SVM algorithms is unsuitable for big data sets due to its high training complexity.	Requires very large amount of data in order to perform better than other techniques
Limitation	World Wide Web contains data in an unstructured format difficult to detect and classify	To detect the multidimensional character of fake news, a few perspectives such as Naïve Bayes classifier, Support Vector Machines, and semantic research must be included.	Social network manipulation is required for attacks on graph-based techniques.

2.3 SUMMARY

According to Table 1, each of the previous studies used different methods, each with its own strengths and weaknesses. However, the most reliable study through my research is Research 2, as it uses two relatively stable methods to detect fake news, produces very high accuracy and resilient behaviour in various difficult environments, and tests the multidimensional characteristics of fake news in real data.

CHAPTER 3

METHODOLOGY

3.1 INTRODUCTION

This chapter introduces how the study carries out to satisfy all of the objectives. It is presented as a flowchart since it is easy to comprehend. It also ensures that the progress is on track.

3.2 RESEARCH FRAMEWORK

This section derived a study methodology for detecting fake news from social media data.

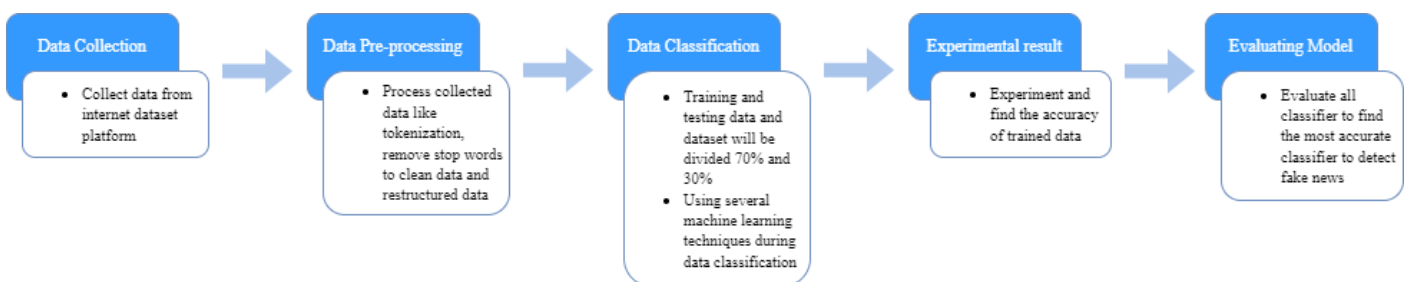


Figure 3.1: Framework of research

3.2.1 Data Collection

Due to the public nature of social media datasets, experimental data for the fake news detection model will be taken from available social media dataset sources on the internet. These data contain text related to the COVID-19 pandemic.

3.2.2 Data Pre-processing

Natural Language Processing (NLP) will be used to process the data. There are few techniques will implement in this process such as tokenization, normalization, stop word removal, stemming, lemmatization. Process of tokenization separates the provided text into smaller parts known as symbols and eliminates any punctuation from the raw textual data. Stop words removal removes unnecessary words such as a, an, about, by, and but. The stemming method is used to extract the fundamental forms of words with the same meaning but distinct word structures, such as connection, connections, connective, and so on. NLP examines the grammatical structure of sentences as well as the specific meanings of words, then use algorithms to extract meaning and provide outputs. In other words, it understands human language so that it can accomplish various activities automatically. Fake news detection favours sentiment analysis to identify the sentiment in the text and categorise news as positive, negative or neutral.

3.2.3 Data Classification

The dataset will be divided into training and testing sections. The training model needs to be pre-processed; 70% of the training data is required depending on the requirements of the experiment. The training data consists of the input and the expected outcome, which contains both the input and the expected output. For the test data, which is used to assess whether the trained model performs effectively on unseen data, 30% of the data is required. Once this data has been thoroughly tested, the chose algorithm is used to predict outcomes.

3.2.4 Experimental Result

Experimental result for fake news detection model is to find the accuracy of trained data. Accuracy of all trained data after run by the machine learning algorithms, the result will be visualized as charts and graphs.

3.2.5 Evaluating Model

Three machine learning algorithms will be used: Support Vector Machine, Naïve Bayes, and Decision Tree. Evaluation model is to evaluate all three algorithms that are selected for training data and test data, based on the accuracy of trained data to find the most accurate classifier to detect fake news on COVID-19.

- Support vector machines (SVM): SVM are machine learning classifiers that may be utilised for both classification and regression tasks (Rushikesh Pupale, 2018). Single kernel SVM are commonly used for data analytics in a variety of domains, notably social media, and linear SVM are regarded as one of the top performing algorithms for text categorization.
- Naïve Bayes (NB): Machine learning model Naïve Bayes is used on classification and regression tasks. It is a classification technique based on Bayes' theorem, which assumes independence between predictors. A plain Bayesian classifier assumes that the presence of a particular feature in a class is independent of the presence of any other feature (Sunil Ray, 2017).
- Decision Tree: The goal of a decision tree is to build a training model for predicting the classes or values of a target attribute by learning basic choice rules from past data (training data) (Nagesh Singh Chauhan, 2022). To predict the class label of a record, the decision tree starts at the root of the tree. The decision tree compares the value of the root attribute with the attributes of the record.

3.3 PROJECT REQUIREMENT

3.3.1 Input

Input of this research given textual data is social media datasets from Zenodo website. In this dataset, the data collection is started from year 2020 and publication date on November 20, 2020. The format of the dataset we will use in this research from Zonodo is csv file format. This dataset is all related to COVID-19 information during the pandemic and will use for fake news detection on COVID-19 in this research paper to find out the feasibility of COVID-19 fake news detection models based on the accuracy on detect fake news.

3.3.2 Output

The accuracy, precision, recall, and F1 of the experimental results from three selected algorithms, Support Vector Machine, Naïve Bayes, and Decision Tree, concerning false news identification, are the outputs of this research. The algorithm below will be used to determine the accuracy of the identification of fake news connected to COVID-19 with four criteria, True Positive, True Negative, False Positive and False Negative.

$$\begin{aligned} \text{Precision} &= \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \\ &= \frac{\text{True Positive}}{\text{Total Predicted Positive}} \end{aligned} \qquad \begin{aligned} \text{Recall} &= \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \\ &= \frac{\text{True Positive}}{\text{Total Actual Positive}} \end{aligned}$$
$$\text{F1} = 2 \times \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy for Binary Classification

Figure 3.2: Accuracy, Precision, Recall, F1 formula

Accuracy is machine learning technique on classifies a data point in correctness. Numerator represents total number of correct detections (TP & TN), while denominator represent total number of detections (TP, TN, FP, FN). The ratio of appropriately classified data to the total number of classifications made by the model is called accuracy (Rodríguez et al., 2022).

Precision refers to how accurate your model is in detecting positives and how many of them are actual positives (Alam et al., 2022). A false positive in fake news detection occurs when a non-fake news item which is actual negative is mistakenly labelled as a fake news item (expected to be a fake news).

Recall measures the real number of positives that the algorithm collects by labelling as true positive(true). In fake news detection, if Actual Positive is detected as non-fake news which is false negative, the result can be very detrimental to fake news detection.

When there is an unequal class distribution, the F1 score may be used to find a balance between Precision and Recall.

3.3.3 Process Description

The process of this research conducted 5 phase which are data collection, data pre-processing, data classification, experimental result, evaluating model. First, we will collect social media datasets from Zenodo website. Second, from the collected datasets, we will do the pre-processing to clean and normalize the data by using tokenization, normalization, stop-words removal, and lemmatization. Next, after the data cleaned and into structured format, we will be classified into 70% training data and 30% test data which training data consists of the input and the expected outcome and test data performs unseen data. Trained data by machine learning algorithms will be used to visualize in graph chart and diagram to do experimental result. Last, we will evaluation on three selected algorithms, Support Vector Machine, Naïve Bayes, and Decision Tree by using the experimental result based on the accuracy of trained data.

3.3.4 Constraints And Limitations

The constraint of this research only three machine learning algorithms will be utilized which is, Support Vector Machine, Naïve Bayes, and Decision Tree. Each algorithm has its own equations that need to be executed, so it will be difficult to implement different machine learning algorithms in this study such as the time required and the amount of data required.

The limitation of this research is that it requires a large dataset to train and test in order to accurately find the accuracy of each algorithm. Second is, the dataset must be processed before it can be trained and tested as a rough dataset is likely to affect the accuracy of the results to proof that which algorithm or model is most compatible for fake news detection.

3.4 PROPOSED DESIGN

3.4.1 Flowchart

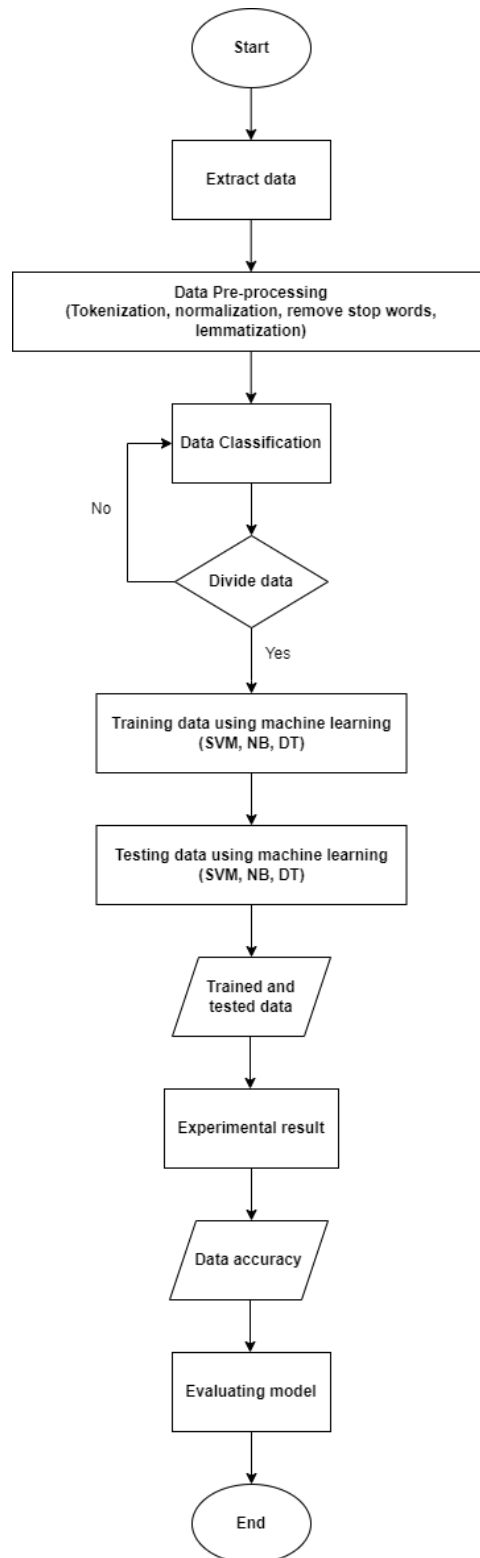


Figure 3.3: Flowchart of research

3.4.2 Flow Explanation

The system starts with data extraction. The data will be taken from available sources on the internet. COVID-19 pandemic related text content and labelling which are 1 equal to true and 0 equal to fake are included in this data.

Next, the data will undergo data pre-processing. This is because the social media text data fetched from the internet may not require special characters, punctuation, numbers and links. The data pre-processing process includes tokenization, stop word deletion, lemmatization and normalisation to give a more uniform format to the data being processed.

The processed data is then sorted into data classification. The data is split into two sections, one for training and one for testing. The training data is being used to run and train machine learning algorithms, while the tested data is used to evaluate the model's accuracy.

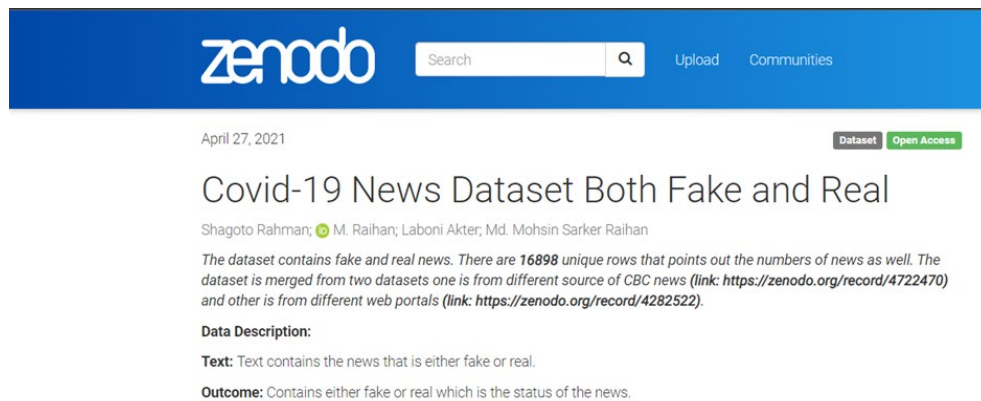
The trained and tested data will be experimented with to discover the accuracy of each algorithm. As an experimental outcome, the accuracy of the trained data will be shown in the form of charts and graphs for comparison.

Lastly, based on the experimental result by generate the graph and chart, the evaluating model will be done by evaluate the three machine learning algorithms used: Support Vector Machine, Naïve Bayes, and Decision Tree to assess which machine learning algorithm runs more accurate training data and is more suitable to be a COVID-19 fake news detection model based on social media data.

3.5 DATA DESIGN

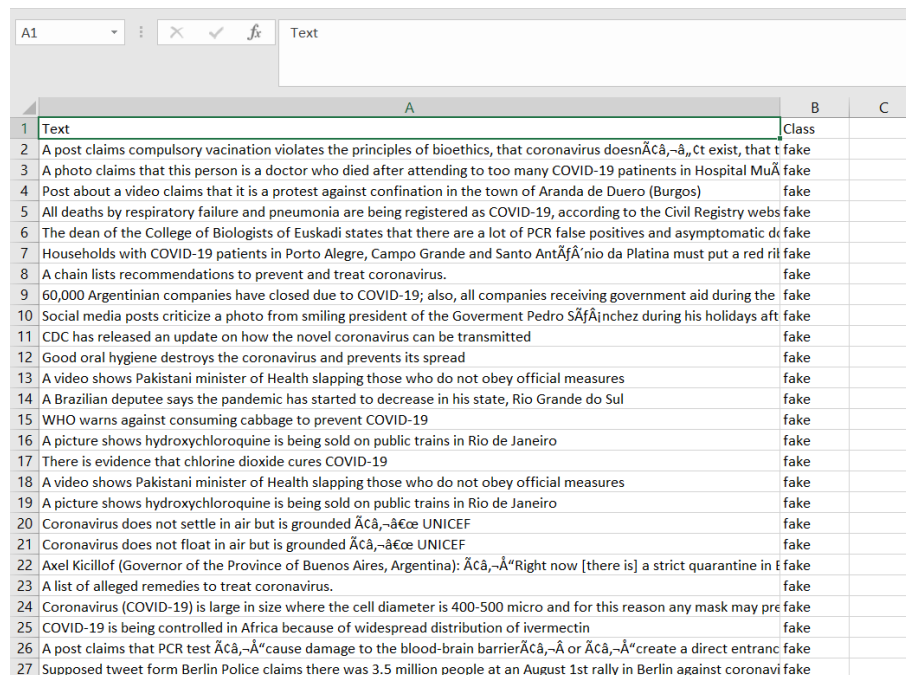
3.5.1 Data Description

The dataset consists of CSV files containing COVID-19 pandemic related text content. The dataset includes fact-checked false news from Poynter as well as true content from news publishers' verified portals. Text, which is user-written text, will be utilised as one of the many data attributes in the dataset. The dataset has been pre-processed, with special characters and non-vital information removed.



The screenshot shows the Zenodo website interface. At the top, there is a search bar and navigation links for 'Upload' and 'Communities'. Below the header, the date 'April 27, 2021' is displayed, along with 'Dataset' and 'Open Access' buttons. The main title is 'Covid-19 News Dataset Both Fake and Real'. The authors listed are Shagoto Rahman, M. Raihan, Laboni Akter, and Md. Mohsin Sarker Raihan. A description states: 'The dataset contains fake and real news. There are 16898 unique rows that points out the numbers of news as well. The dataset is merged from two datasets one is from different source of CBC news (link: <https://zenodo.org/record/4722470>) and other is from different web portals (link: <https://zenodo.org/record/4282522>).' Under 'Data Description', it specifies: 'Text: Text contains the news that is either fake or real.' and 'Outcome: Contains either fake or real which is the status of the news.'

Figure 3.4: Dataset Description



The screenshot shows a spreadsheet with columns A, B, and C. Column A contains news text, column B contains the classification 'fake' or 'real', and column C contains the classification 'Class'. The rows are numbered 1 to 27.

	A	B	C
1	Text		Class
2	A post claims compulsory vaccination violates the principles of bioethics, that coronavirus doesn't exist, that t fake		
3	A photo claims that this person is a doctor who died after attending to too many COVID-19 patients in Hospital MuÃ fake		
4	Post about a video claims that it is a protest against confinement in the town of Aranda de Duero (Burgos) fake		
5	All deaths by respiratory failure and pneumonia are being registered as COVID-19, according to the Civil Registry webs fake		
6	The dean of the College of Biologists of Euskadi states that there are a lot of PCR false positives and asymptomatic di fake		
7	Households with COVID-19 patients in Porto Alegre, Campo Grande and Santo AntÃnio da Platina must put a red ril fake		
8	A chain lists recommendations to prevent and treat coronavirus. fake		
9	60,000 Argentinian companies have closed due to COVID-19; also, all companies receiving government aid during the fake		
10	Social media posts criticize a photo from smiling president of the Government Pedro SÃnchez during his holidays aft fake		
11	CDC has released an update on how the novel coronavirus can be transmitted fake		
12	Good oral hygiene destroys the coronavirus and prevents its spread fake		
13	A video shows Pakistani minister of Health slapping those who do not obey official measures fake		
14	A Brazilian deputeo says the pandemic has started to decrease in his state, Rio Grande do Sul fake		
15	WHO warns against consuming cabbage to prevent COVID-19 fake		
16	A picture shows hydroxychloroquine is being sold on public trains in Rio de Janeiro fake		
17	There is evidence that chlorine dioxide cures COVID-19 fake		
18	A video shows Pakistani minister of Health slapping those who do not obey official measures fake		
19	A picture shows hydroxychloroquine is being sold on public trains in Rio de Janeiro fake		
20	Coronavirus does not settle in air but is grounded Ã UNICEF fake		
21	Coronavirus does not float in air but is grounded Ã UNICEF fake		
22	Axel Kicillof (Governor of the Province of Buenos Aires, Argentina): ÃRight now [there is] a strict quarantine in t fake		
23	A list of alleged remedies to treat coronavirus. fake		
24	Coronavirus (COVID-19) is large in size where the cell diameter is 400-500 micro and for this reason any mask may pre fake		
25	COVID-19 is being controlled in Africa because of widespread distribution of ivermectin fake		
26	A post claims that PCR test Ã cause damage to the blood-brain barrierÃ or Ã create a direct entranc fake		
27	Supposed tweet form Berlin Police claims there was 3.5 million people at an August 1st rally in Berlin against coronavi fake		

Figure 3.5: Table content of dataset

3.6 PROOF OF INITIAL CONCEPT

3.6.1 Machine Learning Algorithms

3.6.1.1 Support Vector Machine

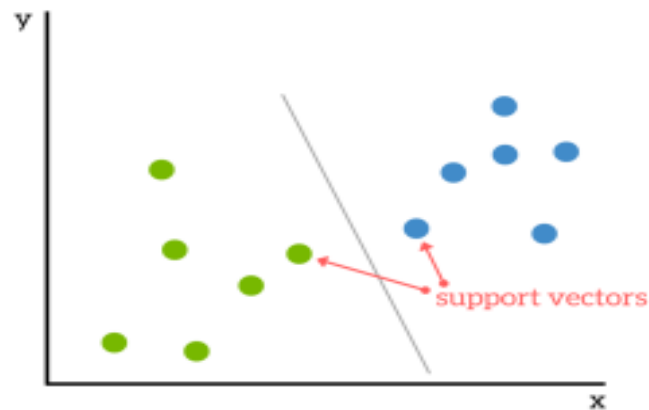


Figure 3.6: Support Vector Machine example

Support vector machines (SVM) are a very well supervised learning-based classification and regression technique for categorising data into several divisions (Alenezi & Alqenaei, 2021). SVM model is to identify an N-dimensional hyperplane that divides the input data into distinct groups. The SVM-based classifier's primary purpose is to discover the appropriate hyperplane with the greatest margin between the numerous hyperplanes that categorise a given data point. It is to discover the hyperplane that best divides the data into two divisions (Jain et al., 2019). Hyperplanes, on the other hand, are decision boundaries that aid machine learning models in classifying data or data points, and the optimal separation occurs when a hyperplane divides two classes by the greatest distance from the nearest data point.

3.6.1.2 Naïve Bayes

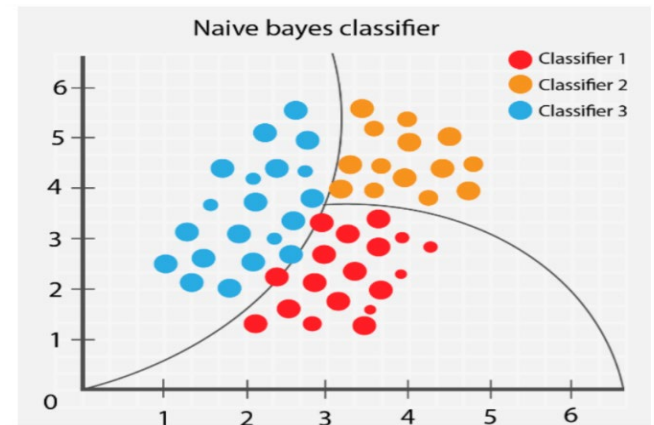


Figure 3.7: Naïve Bayes example

The supervised learning technique based on Bayes' theorem, the Naïve Bayesian algorithm, is used to tackle classification issues. It's typically utilised with huge training datasets for text classification. The Naïve Bayesian classifier is a simple and appropriate classification algorithm that assists in the creation of efficient machine learning algorithms that can make accurate predictions. It's a probabilistic classifier, it produces predictions based on an object's likelihood. One of the most prominent methods for determining the accuracy of news is Naïve Bayes, which utilizes polynomial Naïve Bayes (Jain et al., 2019). The Naïve Bayes classifier connects the usage of tokens (usually words, occasionally other structures, syntactic or not) with false and non-false information, then uses Bayes' theorem to determine the likelihood that a COVID-19 news item is or is not fake news (Yuslee & Abdullah, 2021).

Formula:

$$P(A|B) = \frac{P(A|B)P(A)}{P(B)}$$

Where:

- i. $P(A|B)$ is the Posterior Probability: the probability of hypothesis A for observed event B.
- ii. $P(B|A)$ is the Likelihood Probability: the probability that the hypothesis is true with respect to the evidence.

- iii. $P(A)$ is the Prior Probability: the probability that the hypothesis is true before the evidence is observed.
- iv. $P(B)$ is the Marginal Probability: the probability of the evidence.

3.6.1.3 Decision Tree

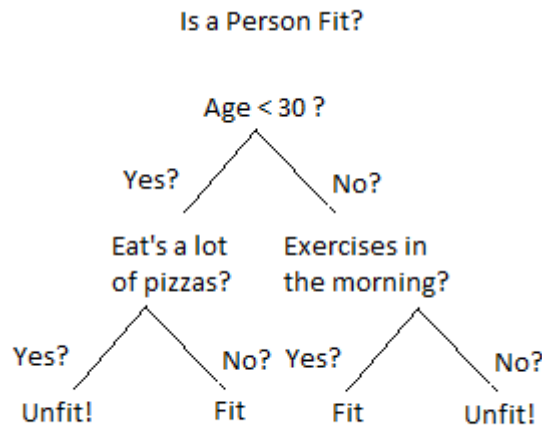


Figure 3.8: Decision Tree example

A decision tree is an advantageous tool that uses a flowchart-like structure to solve classification issues. A decision tree's internal nodes provide a condition or "test" for a feature and branch based on the test condition and the outcome (Khanam et al., 2021). Decision trees are being used in machine learning and data mining to find approximate solutions to data (Somvanshi et al., 2017). Decision trees are utilized to extract data from vast quantities of accessible data using decision rules. Decision trees simply classify data that is easy to store and then classify it again. The decision tree is effective in identifying the most significant factors as well as describing the relationships between them. It is critical in the creation of new variables and properties that are very effective for data exploration and prediction of target variables.

3.7 HARDWARE AND SOFTWARE EQUIPMENT

Below show hardware and software requirement will be utilise in this research paper.

Table 3.1: Hardware equipment

Hardware	Specification	Purpose
Laptop	Katana GF66 11UC CPU: 11th Gen. Intel® Core™ i7 Processor OS: Windows 10 Home	For developing, documenting, and completing research
Smartphone	Huawei P30 Processer: Huawei Kirin 980 Ram: 8GB OS: EMUI 12 (Based on Android)	To aid in finding materials and information needed to complete the project

Table 3.2: Software equipment

Software	Specification	Purpose
Microsoft Office Word	Version 2020	Used for research documentation
Microsoft Office PowerPoint	Version 2020	Used for presentation slide
Draw io	Version 18.0.3	Used to draw flowchart or framework
Google Chrome	Version 102.0.5005.62	Assists in finding previous research papers and materials related to the project
Jupyter Notebook	Python 3.9.13	Process the dataset data and run Python code for research purposes

3.8 POTENTIAL USE OF PROPOSED SOLUTION

Humans are notoriously bad at identifying fake information. This is because bogus news frequently resembles actual news, and we believe we can see a trend. When it comes to spreading information, fake news has an edge. When it comes to sharing information on the internet, individuals seldom analyze it. People are also more prone to spread bad news, and most of the news around COVID-19 is negative.

Through this research, our model can help in COVID-19 fake news detection and provide people with more accurate judgments about COVID-19 fake news. COVID-19 fake news will be implemented through text analysis using python technique. Text analysis is vital for pre-processing and understanding written texts of tweets. Therefore, we will be able to analyse the dataset and detect if it is COVID-19 fake news through text analysis approach. In future study, we will aim to detect fake news using additional machine learning algorithms and experiment with different types of data sets to assess the accuracy and precision of the results.

CHAPTER 4

IMPLEMENTATION, RESULTS AND DISCUSSION

4.1 INTRODUCTION

The purpose of this chapter is to provide explanation of the implementation and result of this research paper. The implementation of fake news detection model by utilised Jupyter Notebook. Social media data on COVID-19. Below in the result of the implementation process using machine learning algorithms.

4.2 RESULT

This research focused on detecting fake news about the COVID-19 using social media data. Data were downloaded from the Zenodo website, which provides a social media dataset on New Crown Pneumonia containing fake news (0) and true news (1). Using this dataset, COVID-19 fake news was estimated and detected in Jupyter Notebook using the python language and different machine learning models. Below is the result obtained by training and testing machine learning models. In order to transform the social media dataset into clean data, the data needs to be pre-processed and cleaned. The data in the dataset from the Zenodo website is pre-processed with tokenization, normalization, stop word removal, stemming, lemmatization to make the dataset clean. The clean news text is then visualised via WordCloud. In WordCloud, the most frequently used words are displayed in a larger font than the less frequently used words. By using WordCloud, the most common words in the dataset will be displayed for both fake news and real news.

In addition, the dataset will be split into 70% training and 30% testing, using different machine learning algorithms (Support Vector Machine, Naïve Bayes and Decision Tree) to train and test the dataset. The results will be calculated using four evaluation metrics: accuracy, precision, recall and f1_score to assess the performance of the machine learning models. In addition, the final metric results will be used to compare which of the different three machine learning algorithms is the most accurate. The results will be displayed in bar charts and graphs. The table below shows the final results of the different metrics using the different machine learning classifiers. The most accurate machine learning algorithm in training and testing is then selected for the implementation of the confusion matrix.

Table 4.1: Accuracy, Precision, Recall and F1_Score of Machine Learning Algorithms

Model	Train				Test			
	Accuracy	Precision	Recall	F1_score	Accuracy	Precision	Recall	F1_score
Support Vector Machine	0.996	0.999	0.991	0.995	0.951	0.959	0.924	0.941
Naïve Bayes	0.959	0.966	0.937	0.951	0.937	0.944	0.906	0.925
Decision Tree	0.999	1	0.999	0.999	0.881	0.853	0.872	0.862

Table 4.1 show the results of Machine Learning Algorithms. From Table 4.1 we can see that the best train Accuracy, 99.9% is achieved by Decision Tree (DT), followed by Support Vector Machine (SVM), 99.6% Accuracy, and the lowest Accuracy 95.9% is generated by Naïve Bayes (NB). While for the best test Accuracy is achieved by Support Vector Machine, 95.1%, followed by Naïve Bayes (NB), 93.7%, and the lowest Accuracy 88.1% is on Decision Tree (DT).

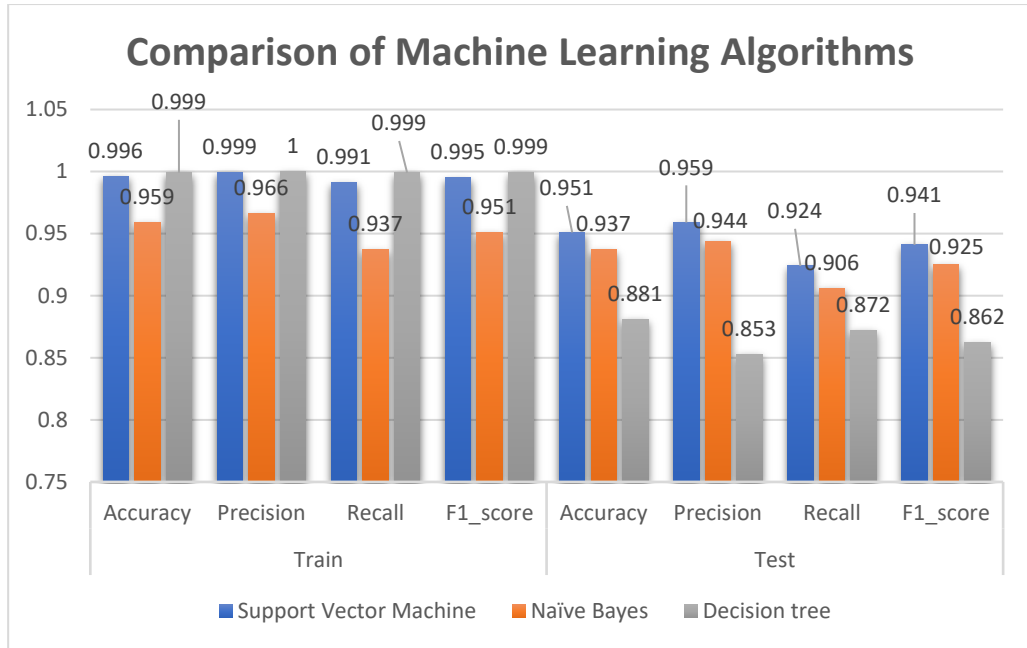


Figure 4.1: Comparison of Machine Learning Algorithms on Accuracy, Precision, Recall, F1_Score

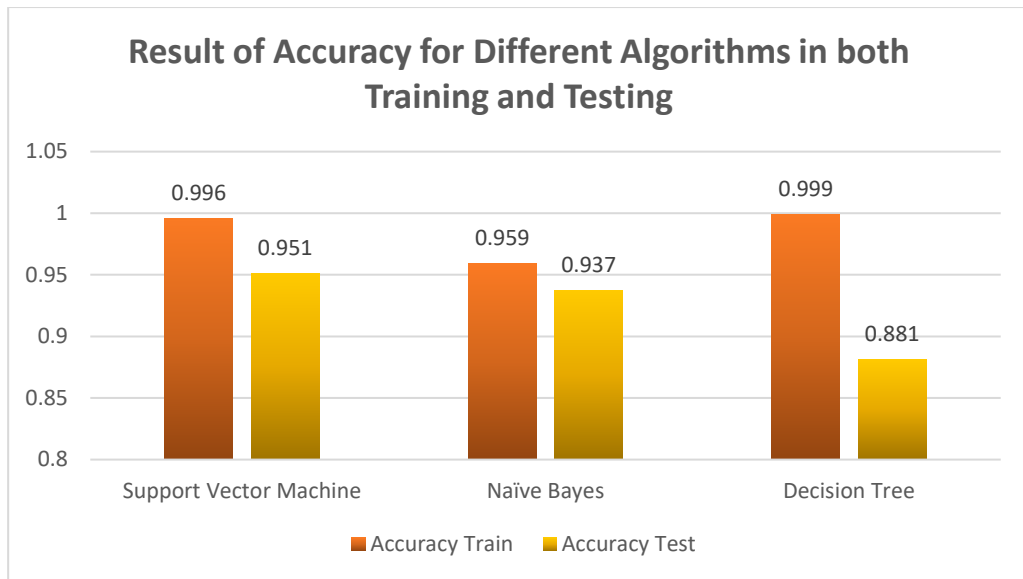


Figure 4.2: Result of Accuracy for Different Algorithms in both Training and Testing

Figure above show that the accuracy results for different algorithms in both training and testing. For training result, the accuracy for Support Vector Machine, Naïve Bayes, and Decision Tree which are 0.996 (99.6%), 0.959 (95.9%), and 0.999 (99.9%). For testing result, the accuracy for Support Vector Machine is 0.951 (95.1%), Naïve Bayes which is 0.937 (93.7%), and Decision Tree which is 0.881 (88.1%).

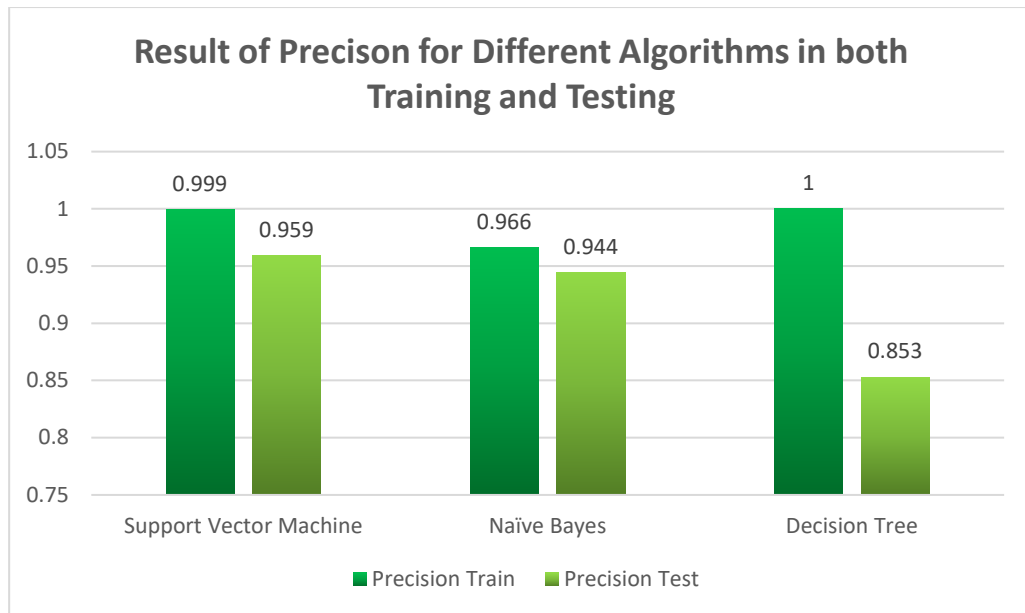


Figure 4.3: Result of Precision for Different Algorithms in both Training and Testing

Figure above show that the precision results for different algorithms in both training and testing. For training result, the precision measure for Support Vector Machine, Naïve Bayes, and Decision Tree which are 0.999 (99.9%), 0.966 (96.6%), and 1 (100%). For testing result, the precision for Support Vector Machine is 0.959 (95.9%), Naïve Bayes which is 0.944 (94.4%), and Decision Tree which is 0.853 (85.3%).

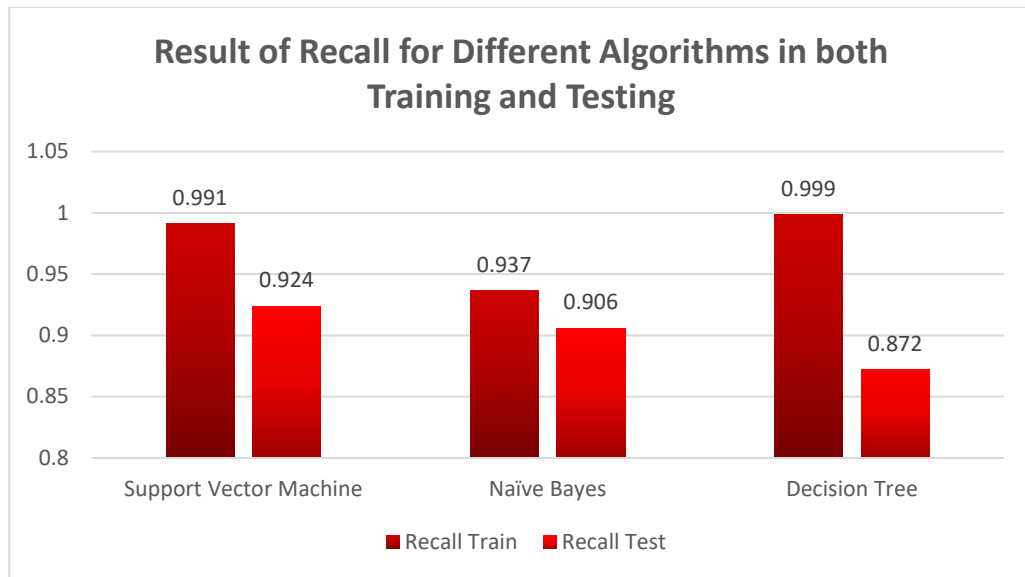


Figure 4.4: Result of Recall for Different Algorithms in both Training and Testing

Figure above show that the recall results for different algorithms in both training and testing. For training result, the precision measure for Support Vector Machine, Naïve Bayes, and Decision Tree which are 0.991 (99.1%), 0.937 (93.7%), and 0.999 (99.9%). For testing result, the precision for Support Vector Machine is 0.924 (92.4%), Naïve Bayes which is 0.906(90.6%), and Decision Tree which is 0.872 (87.2%).

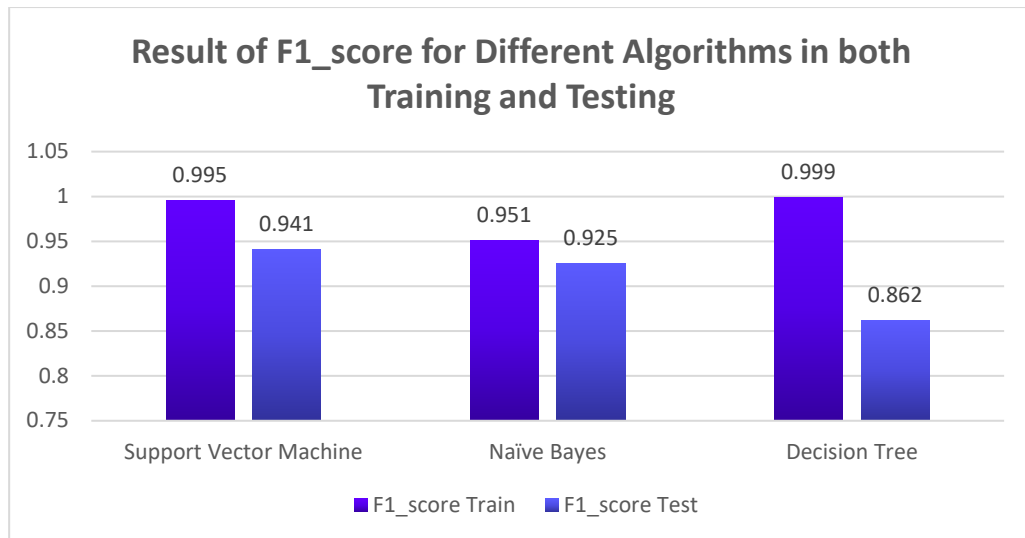


Figure 4.5: Result of F1_score for Different Algorithms in both Training and Testing

Figure above show that the F1_score results for different algorithms in both training and testing. For training result, the precision measure for Support Vector Machine, Naïve Bayes, and Decision Tree which are 0.995 (99.5%), 0.951 (95.1%), and 0.999 (99.9%). For testing result, the precision for Support Vector Machine is 0.941 (94.1%), Naïve Bayes which is 0.925(92.5%), and Decision Tree which is 0.862 (86.2%).

Figures 4.6, 4.7, and 4.8 depict word clouds for fake news, real news, and common words. According to the word clouds and most frequently used terms, there is a large overlap of crucial words between fake and true news. Word clouds will generate all most common words to from the dataset to define positive words (true news) and negative words (fake news).

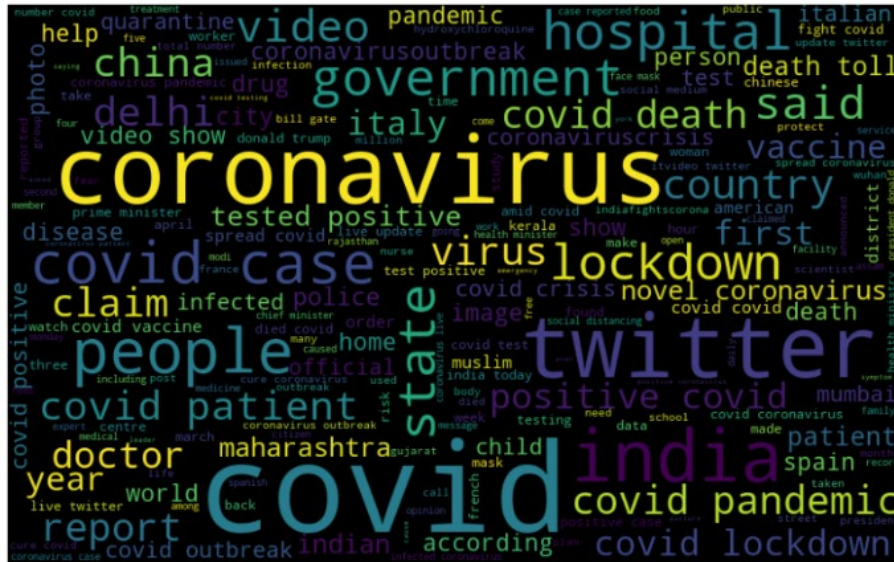


Figure 4.6: Fake and True News common words

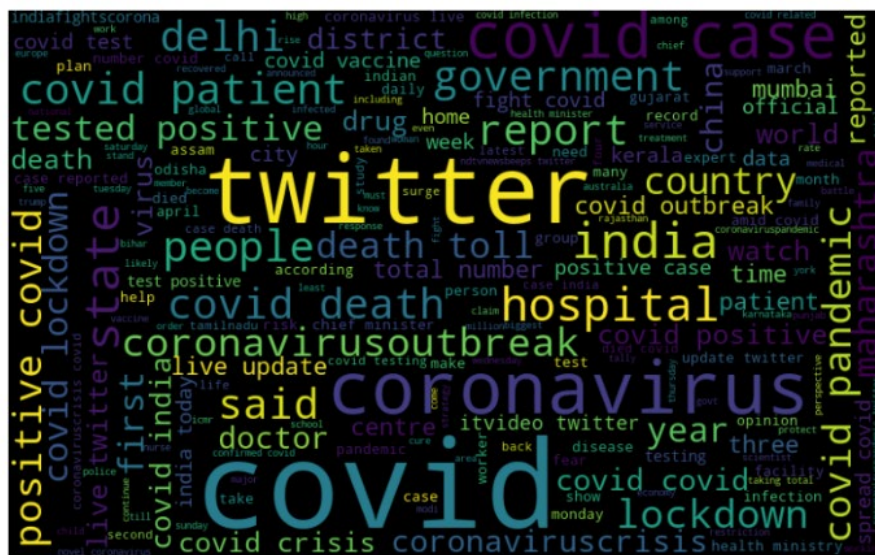


Figure 4.7: Positive words which is true news for the text

The best accuracy for training is the Decision Tree algorithms which have the highest accuracy (99.9%) compared to Support Vector Machine (99.6%) and Naïve Bayes (95.9%). However, for testing data, algorithm with the highest accuracy is Support Vector Machine (95.1%) compared to Naïve Bayes (93.7%) and Decision Tree (88.1%). Below is the figure of confusion matrix based on the best accuracy for training and testing data.

Figure 4.9, 4.10 show the best confusion matrix of covid-19 fake news detection on the training and test data. In Figure 4.9 is confusion matrix of Decision Tree (DT) and 4.10 is confusion matrix of Support Vector Machine (SVM). 1 is represent true news and 0 is represent fake news. In training data for Decision Tree algorithm confusion matrix, 6809 results out of 6809 is true negative which is fake news and 5082 results out of 5083 is true positive which is true news. While the test data for Support Vector Machine confusion matrix. 2832 results out of 2918 is true negative which is fake news and 2014 result out of 2179 is true positive which is true new.

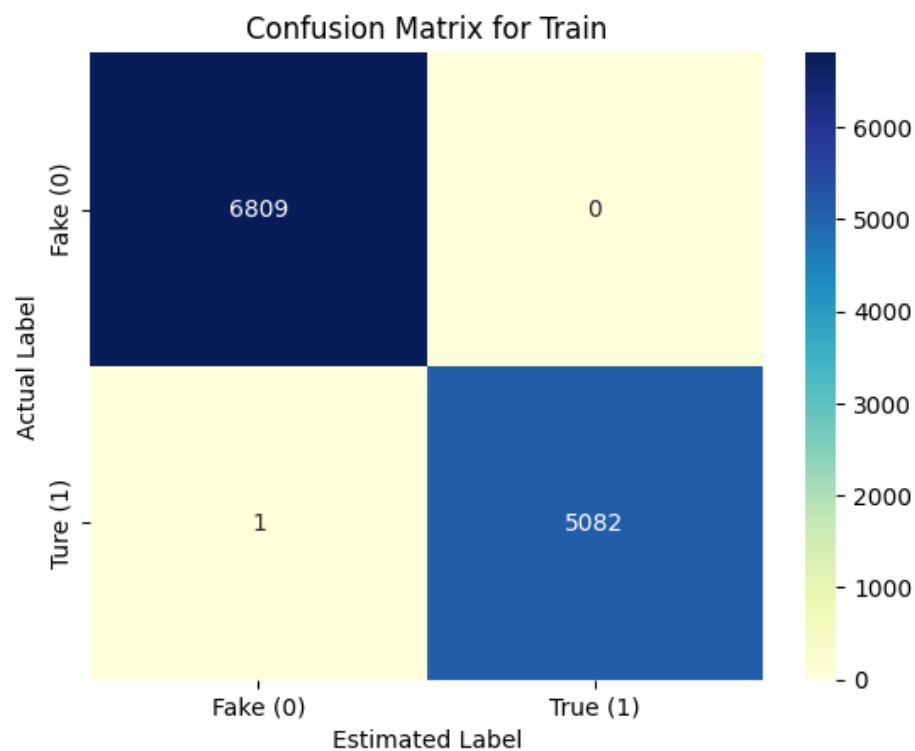


Figure 4.9: Confusion matrix of Decision Tree in Training

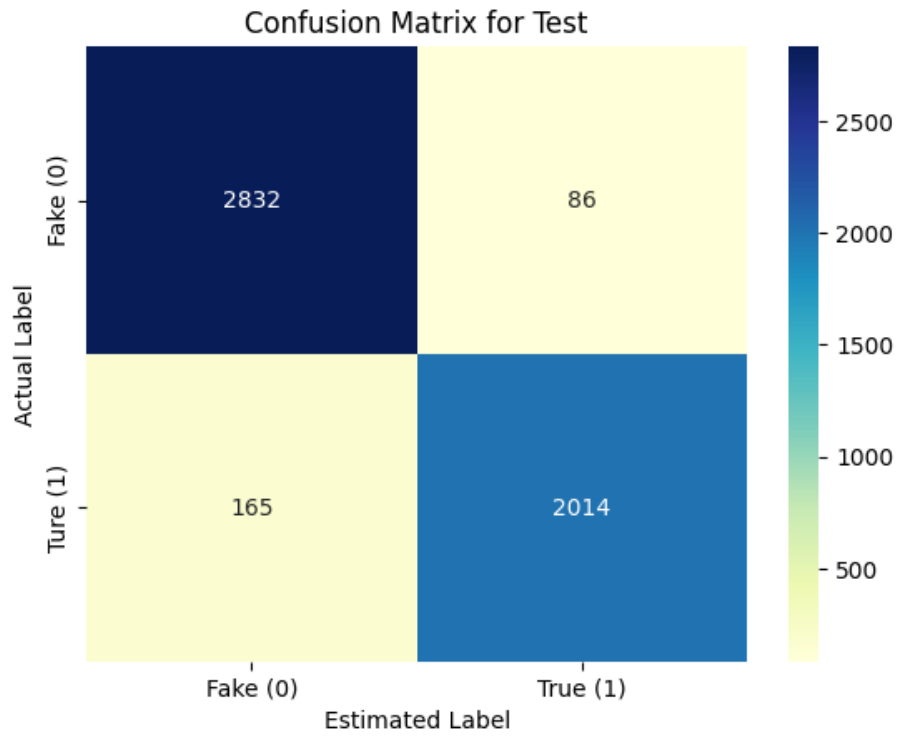


Figure 4.10: Confusion matrix of Support Vector Machine in Testing

4.3 DISCUSSION

For conclusion, Decision Tree algorithms have the best performance in the training data out of three algorithms. However, for the testing data, Support Vector Machine has achieved the highest accuracy performance. Therefore, this study was implemented to study the attributes of texts related to COVID-19 fake news through social media data, using the content of the dataset of which is text related to COVID-19 fake news and to test the accuracy of social media news about COVID-19 by using machine learning techniques to detect fake news about COVID-19.

CHAPTER 5

CONCLUSION

5.1 OBJECTIVE REVISITED

For this thesis, the objective of this research has been stated which is to study the use of social media data for Covid-19 fake news identification. Throughout the process by using machine learning algorithm and natural language processing on training and testing social media data, accuracy of the classifier support vector machine (SVM) is considered to be the most suitable for designing the Covid-19 fake news detection model. By using the developed Covid-19 fake news detection model, user able to detect the credibility of Covid-19 news on any social media whether it is true or fake.

5.2 LIMITATION

Limitation of a study are its flaws and shortcomings that impacted or influenced the interpretation of the findings from your research. Firstly, is the small data sample size. While there are stories about this coronavirus news around the globe. But because the Covid-19 emerged in recent years, the data integration of its news or information on social media is not quite much. For this study, the data set came from Zenodo, an open repository platform. The data set size used to train and test the algorithms is only 16989. The more data there is, the more accurately the algorithm can be trained and tested. Due to the small sample size of data, the accuracy of each algorithm is generally high in training.

Next is the implementation of data collection method. Because we do not have an extensive experience in primary data collection, there is a great chance that the nature of implementation of data collection method is flawed. We only able to use the dataset that already provided on the dataset publishing platform rather than collecting the data from social media directly. Data collected in the data set publishing platform may not have been updated for a long time and therefore may have no reference value.

5.3 FUTURE WORK

There are some limitations to this subject, but there are also some improvements for future research. Due to the COVID-19 expansion around the world, there will be more information and news in the future about the COVID-19. In the future, a larger amount of data can be obtained for training and testing, so as to test the accuracy of each algorithm for detecting fake news of COVID-19. We also hope to expand this type of research into other areas, not just to detect fake news about COVID-19, but to detect other types of fake news. Secondly, we should experiment with more classifiers and algorithms. For example, Logistic Regression, Random Forests, Recurrent Neural Network, Neural Network, K-Nearest Neighbour, etc. For detecting fake news, SVM and Naive Bayes classifiers appear to be the best, according to experiments and studies conducted by researchers. Since no algorithm is the best, trying other algorithms will make the studies more comparable and not limited to a single algorithm. We look forward to investigating the performance of the new classification method, which will improve the accuracy of the COVID-19 fake news detection model.

REFERENCE

1. Agrawal, C., Pandey, A., & Goyal, S. (2021). A Survey on Role of Machine Learning and NLP in Fake News Detection on Social Media. *2021 IEEE 4th International Conference on Computing, Power and Communication Technologies, GUCON 2021*, 1–7. <https://doi.org/10.1109/GUCON50781.2021.9573875>
2. Alam, A. M., Kurum, M., & Gurbuz, A. C. (2022). Radio Frequency Interference Detection for SMAP Radiometer Using Convolutional Neural Networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, PP(DI)*, 1–15. <https://doi.org/10.1109/JSTARS.2022.3223198>
3. Alenezi, M. N., & Alqenaei, Z. M. (2021). Machine learning in detecting covid-19 misinformation on twitter. *Future Internet, 13*(10), 1–20. <https://doi.org/10.3390/fi13100244>
4. Jain, A., Shakya, A., Khatter, H., & Gupta, A. K. (2019). A smart System for Fake News Detection Using Machine Learning. *IEEE International Conference on Issues and Challenges in Intelligent Computing Techniques, ICICT 2019*, 3–6. <https://doi.org/10.1109/ICICT46931.2019.8977659>
5. Khanam, Z., Alwasel, B. N., Sirafi, H., & Rashid, M. (2021). Fake News Detection Using Machine Learning Approaches. *IOP Conference Series: Materials Science and Engineering, 1099*(1), 012040. <https://doi.org/10.1088/1757-899x/1099/1/012040>
6. Monti, F., Frasca, F., Eynard, D., Mannion, D., & Bronstein, M. M. (2019). *Fake News Detection on Social Media using Geometric Deep Learning*. 1–15. <http://arxiv.org/abs/1902.06673>
7. Nagesh Singh Chauhan. (2022, February 9). *Decision Tree Algorithm, Explained - KDnuggets*. KDnuggets. <https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html>
8. Oshikawa, R., Qian, J., & Wang, W. Y. (2020). A survey on natural language processing for fake news detection. *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, 6086–6093.
9. Rodríguez, M., Alesanco, Á., Mehavilla, L., & García, J. (2022). Evaluation of Machine Learning Techniques for Traffic Flow-Based Intrusion Detection.

Sensors, 22(23). <https://doi.org/10.3390/s22239326>

10. Rushikesh Pupale. (2018, June 16). *Support Vector Machines(SVM) — An Overview | by Rushikesh Pupale | Towards Data Science*. Towards Data Science. <https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989>
11. Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). *Fake News Detection on Social Media: A Data Mining Perspective*. i. <http://arxiv.org/abs/1708.01967>
12. Somvanshi, M., Chavan, P., Tambade, S., & Shinde, S. V. (2017). A review of machine learning techniques using decision tree and support vector machine. *Proceedings - 2nd International Conference on Computing, Communication, Control and Automation, ICCUBEA 2016*. <https://doi.org/10.1109/ICCUBEA.2016.7860040>
13. Sunil Ray. (2017, September 11). *Learn Naive Bayes Algorithm | Naive Bayes Classifier Examples*. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/>
14. Villagracia Octaviano, M. (2021). Fake News Detection Using Machine Learning. *ACM International Conference Proceeding Series, 2020*, 177–180. <https://doi.org/10.1145/3485768.3485774>
15. Yuslee, N. S., & Abdullah, N. A. S. (2021). Fake News Detection using Naive Bayes. *2021 IEEE 11th International Conference on System Engineering and Technology, ICSET 2021 - Proceedings*, 112–117. <https://doi.org/10.1109/ICSET53708.2021.9612540>

APPENDIX A

Gantt Chart

Task and Milestones	Month						
	March 2022	April 2022	May 2022	Jun 2022	Nov 2022	Dec 2022	Jan 2023
First meeting with supervisor							
Chapter 1 - Introduction							
Chapter 2 – Literature review							
Chapter 3 - Methodology							
PSM 1 submission							
Implementation fake news detection model							
Train and test data for model							

Chapter 4 - Result and Discussion							
Chapter 5 - Conclusion							