# VACCINE HESITANCY DETECTION USING BERT FOR MULTIPLE SOCIAL MEDIA PLATFORMS

BY

SHEIKH MD HANIF HOSSAIN

A thesis submitted in fulfillment of the requirement for the degree of Master of Computing (Computer Science and Information Technology).

Kulliyyah of Information and Communication Technology
International Islamic University Malaysia

OCTOBER 2023

# ABSTRACT

Vaccination has been proven to be an effective measure to prevent the spread of harmful diseases. Despite its efficacy, the moves towards vaccine hesitancy have been receiving global attention. Vaccine hesitancy issues have been openly discussed across major social media platforms including Facebook, Reddit, Twitter, Instagram and YouTube. The spread of vaccine hesitancy-related posts is propagated substantially, causing greater threats to public health. Consequently, various state-of-the-art machine learning techniques have been proposed to analyse vaccine-hesitant related posts in social media. One of the most recent approaches is the transfer learning method using a pre-trained Bidirectional Encoder Representations from Transformers (BERT) model. Despite vaccine hesitancy being a prevalent issue across multiple social media platforms, only a few studies have utilised data from multiple social media platforms to detect vaccine hesitancy. To address this research gap, the use of BERT as one of the new language representation models is adopted to train from a collection of vaccine hesitancy related data from multiple social media platforms. Moreover, this study employs the Support Vector Machine (SVM) and Logistic Regression (LR) models and compare their performances against the BERT method. The objectives of this research are threefold; to establish a consolidated dataset from multiple social media sources for use in vaccine hesitancy detection, to evaluate the effectiveness of using mono-platform versus multi-platform vaccine hesitancy data on the performance of different machine learning models and to apply a transfer learning method using BERT in vaccine hesitancy detection. A collection of 193,023 labelled vaccine hesitant posts were aggregated from three (3) social media platforms which includes Facebook, Reddit, and Twitter. The results demonstrate that the BERT model performs the best and achieved an F1-score of 0.93, while both the SVM and LR achieved F1-scores of 0.90 when detecting vaccine hesitancy from multiple social media platforms. Our proposed research also revealed that models trained with multi-platform data perform at least 15% better than models trained with mono-platform data when tested with multi-platform data.

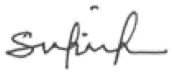
# خلاصة البحث

ثبت أن التطعيم هو إجراء فعال لمنع انتشار الأمراض الضارة. على الرغم من فعاليتها، إلا أن نحو تردد اللقاحات قد حظيت باهتمام عالمي. تمت مناقشة قضايا تردد اللقاحات التحركات و Facebook منصات وسائل التواصل الاجتماعي الرئيسية بما في ذلك بشكل علني عبر تم انتشار المشاركات المتعلقة .YouTube و Instagram و Twitter و Reddit على ذلك، يتسبب في تهديدات أكبر للصحة العامة. وبناءً بالتردد في اللقاح منتشر بشكلكبير، مما لتحليل المنشورات ذات الصلة باللقاحات المترددة تم اقتراح العديد من تقنيات التعلم الآلي الحديثة الأساليب هي طريقة تعلم النقل في وسائل التواصل الاجتماعي.كانت واحدة من أحدث على .(BERT) باستخدام تمثيلات التشفير ثنائية الاتجاه المدربة مسبقًا من نموذج المحولات الاجتماعي، الرغم منكون تردد اللقاح مشكلة منتشرة عبر العديد من منصات وسائل التواصل متعددة فقط من الدراسات قد استخدمت بيانات من منصات وسائط اجتماعية إلا أن عدًدا قليلاً كأحد نماذجBERT لاكتشاف تردد اللقاح. لمعالجة هذه الفجوة البحثية، تم اعتماد استخدام تمثيل اللغة الجديدة للتدريب من مجموعة من البيانات المتعلقة بالتردد في اللقاح من منصات وسائط (SVM) اجتماعية متعددة. علاوة على ذلك، تستخدم هذه الدراسة نماذج آلة المتجهات الداعمة أهداف هذا البحث ثلاثية. .BERT ومقارنة أدائها مع طريقة (LR) والانحدار اللوجستي مجموعة بيانات موحدة من مصادر وسائط اجتماعية متعددة لاستخدامها في الكشف لإنشاء لتحليل تأثير استخدام بيانات تردد اللقاح الخاصة بمنصة محددة مقابل منصات عن تردد اللقاح، BERT التعلم الآلي المختلفة ولتقييم فعالية طريقة تعلم النقل استخدام متعددة على أداء نماذج تجميع مجموعة من 193023 منشورا متردًدا بشأن اللقاح من في الكشف عن تردد اللقاح. تم .Twitter ًوReddit و Facebook تشمل ثلاث (3) منصات وسائط اجتماعية قدرها 0.93، بينما F1 أداء وحقق درجة يحقق أفضل BERT توضح النتائج أن نموذج عند اكتشاف تردد اللقاح من منصات عند 0.90 F1 درجات LR و SVM حققكل من

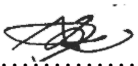
المدربة على بيانات متعددة ًضا أن النماذج وسائط اجتماعية متعددة.كشف بحثنا المقترح أي أفضل بنسبة 15٪ على الأقل من النماذج المدربة ببيانات المنصات تؤدي أداءً خاصة بالمنصة عند اختبارها باستخدام بيانات متعددة المنصات.

# APPROVAL PAGE

I certify that I have supervised and read this study and that in my opinion, it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a thesis for the degree of Master of Computing (Computer Science and Information Technology)



…….
Suriani Sulaiman
Supervisor



…….
Norlia Md Yusof
Co-Supervisor

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a thesis for the degree of Master of Computing (Computer Science and Information Technology)

…….
Amelia Ritahani Ismail
Internal Examiner

…….
Nurulhuda Firdaus Bt Mohd azmi
External Examiner (If Applicable)

This thesis was submitted to the Department of Computer Science and is accepted as a fulfilment of the requirement for the degree of Master of Computing (Computer Science and Information Technology)

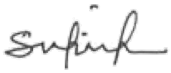
…….
Amir Aatieff Bin Amir Hussin
Head, Department of Computer Science

This thesis was submitted to the Kulliyyah of Information and Communication Technology and is accepted as a fulfillment of the requirement for the degree of Master of Computing (Computer Science and Information Technology)

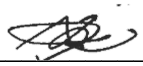
…….
Murni Mahmud,
Dean, Kulliyyah of Information and Communication Technology

# APPROVAL PAGE

The thesis of Sheikh Md Hanif Hossain has been approved by the following:



Suriani Sulaiman
Supervisor



Norlia Md Yusof
Co-supervisor

Amelia Ritahani Ismail
Internal Examiner

Nurulhuda Firdaus Bt Mohd azmi
External Examiner

Zahidah Binti Zulkifli
Chairman

# DECLARATION

I hereby declare that this dissertation is the result of my own investigations, except where otherwise stated. I also declare that it has not been previously or concurrently submitted as a whole for any other degrees at IIUM or other institutions.

Sheikh Md Hanif Hossain

Signature……  Date…16/10/23……

**INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA**

**DECLARATION OF COPYRIGHT AND AFFIRMATION OF FAIR USE OF UNPUBLISHED RESEARCH**

**VACCINE HESITANCY DETECTION USING BERT FOR MULTIPLE SOCIAL MEDIA PLATFORMS**

I declare that the copyright holder of this thesis/dissertation are jointly owned by the student and IIUM.



No part of this unpublished research may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise without prior written permission of the copyright holder except as provided below

1. Any material contained in or derived from this unpublished research may only be used by others in their writing with due acknowledgement.

2. IIUM or its library will have the right to make and transmit copies (print or electronic) for institutional and academic purpose.

3. The IIUM library will have the right to make, store in a retrieval system and supply copies of this unpublished research if requested by other universities and research libraries.

By signing this form, I acknowledged that I have read and understand the IIUM Intellectual Property Right and Commercialization policy.

Affirmed by Sheikh Md Hanif Hossain

…… 16/10/23 ……

Signature Date

**INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA**

**DECLARATION OF COPYRIGHT AND AFFIRMATION OF FAIR USE OF UNPUBLISHED RESEARCH**

**VACCINE HESITANCY DETECTION USING BERT FOR MULTIPLE SOCIAL MEDIA PLATFORMS**

I declare that the copyright holder of this thesis/dissertation is International Islamic University Malaysia.





Affirmed by Sheikh Md Hanif Hossain

…… …16/10/23..

Signature Date

**INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA**

**DECLARATION OF COPYRIGHT AND AFFIRMATION OF FAIR USE OF UNPUBLISHED RESEARCH**

**VACCINE HESITANCY DETECTION USING BERT FOR MULTIPLE SOCIAL MEDIA PLATFORMS**

I declare that the copyright holder of this thesis/dissertation is Sheikh Md Hanif Hossain.





Affirmed by Sheikh Md Hanif Hossain

…… 16/10/23

Signature Date

## NO TITLE FOR THIS PAGE

*This thesis is dedicated to my parents for laying the foundation of what I turned out to be in life.*

# ACKNOWLEDGEMENTS

All credit and commendation should be directed solely to Allah, whose generosity and compassion have been with me during my academic journey. Despite the obstacles I encountered, the blessings and grace of Allah enabled me to surmount the challenges and successfully complete my thesis.

I would like to extend my sincere gratitude to everyone who has encouraged and assisted me as I have finished this thesis. Foremost, I am grateful to my thesis supervisor, Dr. Suriani Bt Sulaiman, for her important advice, criticism, and support during the research process. Without her assistance the completion of the thesis wouldn't be possible. Additionally, I would like to extend my thanks to my co-advisor, Dr. Norlia Md Yusof, for her insightful feedback.

I also want to express my gratitude to my family, for their everlasting love and support. Their confidence in me and never-ending support were crucial to my academic development.

And finally, I owe a debt of gratitude to the academic community and all the researchers who have made contributions to my field of study. Their work paved the road for mine and motivated me to carry on learning and make a difference.

Once again, I express my gratitude to Allah for His infinite mercy, including the fact that He enabled me to complete this thesis successfully. Alhamdulillah.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| BERT | Bidirectional Encoder Representations from Transformers NLP Natural Language Processing |
| BiLSTM | Bidirectional Long Short-term Memory CNN Convolutional Neural Network |
| CSV | comma-separated value |
| DNN | Dense Neural Network |
| HPV | Human Papillomavirus |
| LASSO | Least Absolute Shrinkage and Selection Operator |
| LDA | Latent Dirichlet Allocation |
| LinearSVC | Linear Support Vector Classification |
| LR | Logistic Regression |
| LSTM-LM | Long Short-term Memory Language Model API Application Programming Interface |
| MMR | Measles, Mumps, and Rubella |
| NB | Naïve Bayes |
| NN | Neural Network |
| RDF | Radial Basis Function |
| RF | Random Forest |
| ROC | Receiver Operating Characteristics |
| SMOTE | Synthetic Minority Oversampling Technique |
| SV | Sentence Vector |
| SVM | Support Vector Machine |
| SVM | Support Vector Machine |
| VADER | Valence Aware Dictionary and sEntiment Reasoner TF-IDF Term Frequency-Inversed Document Frequency |
| WHO | World Health Organisation |

# CHAPTER ONE
# INTRODUCTION

## 1.1 BACKGROUND OF THE STUDY

The impact of immunizations on human well-being and lifespan is considered one of the most noteworthy segments in scientific history. Many places in the world have seen significant improvements in health since 1950. Over the last two decades, the average life expectancy has increased while the global new-born death rate has dropped by two-thirds (Plotkin, 2014). Smallpox and polio have both been declared eradicated (Bloom, 2011). However, the spread of vaccine hesitancy moves are of greater threats to public health.

The World Health Organisation (WHO) defines vaccine hesitancy as a "delay in acceptance or refusal of vaccines despite availability of vaccination services". In this era of social media, vaccine misinformation is spreading at an accelerated rate. As a result, more people are openly expressing their opposition towards vaccination (Muric,Wu, & Ferrara, 2021). Researchers are trying to detect such an outbreak by using various analytical and cutting-edge approaches. One of the most popular approaches is to implement machine learning or deep learning techniques to identify vaccine hesitancy from social media (Yuan & Crooks, 2018; Yoo et al., 2019).

Various machine learning techniques have been utilised to detect vaccine hesitancy on social media platforms. One of the most popular methods is Bidirectional Encoder Representations from Transformers (BERT) which is a transformer-based machine learning method developed by Google (Zhang et al., 2020; Liu et al., 2021; Lemmens et al., 2021). BERT has already been trained by using huge amounts of data from Wikipedia and book corpus in which re-training of the model is not required. Thus, to perform natural language processing related tasks such as sentiment analysis, question answering and language inference, BERT only needs to be trained and fine-tuned on small amounts of datasets and additional output layers without the need for task specific architectures (Devlin

et al., 2019). Its unified architecture across diverse tasks saves computational expenses of having to train the whole model. In a recent study by Hossain and Sulaiman (2022), it is also found that the performance of the BERT model for vaccine hesitancy detection on a single social media platform was the second best after Support Vector Machine (SVM).

## 1.2 STATEMENT OF THE PROBLEM

Vaccine hesitancy is a trending issue that is being discussed across numerous online social media platforms, including Facebook, Twitter, Instagram, and others (Garay et al., 2019; Argyris et al., 2021; Wang et al., 2021; Meppelink et al., 2021). To automatically detect anti-vaccine sentiments, various studies using machine learning approaches have been conducted that largely focused on a single social media platform. Only a handful of researchers incorporated data from multiple social media platforms for model training (Lemmens et al., 2021). This singular concentration on a single platform is problematic since there is no guarantee that these models will generalise well across platforms (Salminen et al., 2020). In other words, a model developed by using data from a single social media platform may not perform well in detecting vaccine hesitancy on other platforms. The mono-platform approach is also less efficient since the lack of a universal vaccine hesitancy model forces scholars and scientists to "reinvent the wheel", which means that a new classifier must be constructed each time vaccine hesitancy research is conducted on a specific social media platform. This results in a waste of intellectual effort since repeated studies are conducted using the same techniques across multiple social media platforms. In general, platform specific approaches create unnecessary difficulties for vaccine hesitancy detection across platforms. Thus, there is a need for a vaccine hesitancy classifier built with multi-platform data that performs well across multiple social media platforms.

## 1.3 RESEARCH OBJECTIVES

The objectives of this research are as follows:

i. To establish a consolidated dataset from multiple social media sources for use in vaccine hesitancy detection.
ii. To evaluate the effectiveness of using mono-platform versus multi-platform vaccine hesitancy data on the performance of different machine learning models.
iii. To apply a transfer learning method using BERT in vaccine hesitancy detection.

## 1.4 RESEARCH QUESTIONS

Our research questions are as follows:

i. RQ1: What are the datasets available for use in vaccine hesitancy detection?
ii. RQ2: How can multi-platform data improve the accuracy of vaccine hesitancy detection among different machine learning models?
iii. RQ3: How well does the BERT-based transfer learning model perform in the detection of vaccine hesitancy on multiple social media platforms?

## 1.5 SIGNIFICANCE OF THE RESEARCH

The present research attempts to make a valuable contribution to the field of vaccine hesitancy detection by developing a multi-platform vaccine hesitancy classifier. This classifier is expected to perform effectively and generalise well across a wide range of social media platforms. To this end, the study aims to create a consolidated dataset that can serve as a valuable resource for other researchers who may wish to explore and develop more efficient methods for vaccine hesitancy detection. In addition to its potential use by fellow researchers, the proposed multi-platform vaccine hesitancy classifier may also be utilised by health practitioners to monitor public sentiment towards vaccination on social media platforms.

## 1.6 SCOPE OF THE RESEARCH

The current study aims to create a consolidated dataset comprising social media data from Facebook, Reddit, and Twitter only. Specifically, the dataset will consist of text-based postings in the English language exclusively. To prepare the dataset for training the BERT model, graphical-based emojis will be excluded. The models used in this study are limited to BERT-base, BERT-small, SVM and Logistic Regression (LR). It is worth noting that the research does not focus on identifying vaccine hesitancy associated with a particular vaccine type; rather, it seeks to analyse the sentiment towards diverse types of vaccines.

# CHAPTER TWO
# LITERATURE REVIEW

## 2.1 INTRODUCTION

In Chapter 1, the background of the research, research objectives, research questions and the scope of the research have been introduced. The chapter outlined here encompasses a thorough exploration of vaccine hesitancy detection techniques, underpinned by a robust theoretical background. Section 2.2 focuses on the theoretical foundation, Section 2.3 delves into various vaccine hesitancy detection methods and section 2.4 provides the summary of the chapter.

In Section 2.2, the theoretical background, the reader is provided with a comprehensive definition of vaccine hesitancy (Section 2.2.1). Furthermore, the chapter elucidates machine learning principles, with a particular emphasis on two widely used algorithms, Logistic Regression (LR) and Support Vector Machine (SVM) (Section 2.2.2). It also offers a succinct overview of deep learning (Section 2.2.3). Section 2.2.4 delves into transfer learning, specifically highlighting the BERT model. Section 2.2.5 explores Natural Language Processing (NLP) and sentiment analysis

Moving on to Section 2.3, which delves into vaccine hesitancy detection techniques are discussed and is further divided into four subsections. These include analytical techniques (Section 2.3.1), machine learning techniques (Section 2.3.2), deep learning and ensemble learning techniques (Section 2.3.3), and transfer learning techniques (Section 2.3.4). These subsections provide a comprehensive overview of the literature in each respective area, offering insights into the methodologies employed in detecting vaccine hesitancy from social media. Section 2.3.5 ends Section 2.3 by presenting a detailed analysis of the research gaps within the area of vaccine hesitancy detection techniques.

## 2.2 THEORETICAL BACKGROUND

### 2.2.1 Vaccine Hesitancy

Vaccine hesitancy refers to a delay in accepting or refusing immunisation despite readily available vaccination services (MacDonald, 2015). The '5C model of the drivers of vaccine hesitancy' is a framework that was created based on research conducted in developed nations. It outlines five key factors at the individual level that contribute to vaccine hesitancy: complacency, confidence, convenience (or barriers), collective responsibility and risk assessment (Betsch et al., 2018). Social media is one of the most effective ways to disseminate vaccination misinformation online which creates vaccine hesitancy among the public (Renee Garett et al., 2021). It has become a major source of information for people around the world. However, it is also a breeding ground for misinformation and disinformation, which can contribute to vaccine hesitancy (Skafle et al., 2022). This misinformation can take many forms, such as false claims about the safety or efficacy of vaccines, or conspiracy theories about the motives of vaccine manufacturers (Cerda & García, 2021). A report by the Centre for Countering Digital Hate (CCDH) in the United States indicates that social media accounts belonging to anti-vaccination proponents have amassed 7.8 million followers since 2019. The report also reveals that 31 million Facebook users follow anti-vaccination accounts, while 17 million YouTube users subscribe to similar accounts (Burki et al., 2020). Anti-vaccination clusters were shown to have a much higher presence and were strongly entwined with unsure clusters in a study of vaccine talk on Facebook (Johnson et al., 2020). According to Wilson et al.(2020), using social media to organise action is linked to the notion that vaccines are dangerous, and foreign disinformation tactics on social media are linked to decreased vaccination rates. Vaccine hesitancy can have serious consequences for public health. It can lead to outbreaks of preventable diseases, such as measles and mumps. It can also make it more difficult to achieve herd immunity, which is essential for protecting vulnerable populations from disease (Jafar et al., 2021).

### 2.2.2 Machine Learning

According to Tom Mitchell (1997), "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance a task in T, as measured by P, improves with experience E." In other words, machine learning involves developing computer programs that can perform complex tasks and make predictions based on historical data, which is one of its primary strengths.

There are mainly four categories of machine learning (ML) approaches: supervised, unsupervised, semi-supervised and reinforcement learning. In supervised ML, the final output of the dataset is labelled or known which requires human expertise to label the output of the dataset. On the contrary, unsupervised ML does not require labelling the data. It is used when the output of the input features is uncertain. Data labelling is a costly and a time-consuming process. Hence, semi-supervised ML is used with some labelled and unlabelled data. Finally, reinforcement learning is where an agent is rewarded (positively or negatively) based on the action chosen. The model is then modified in line with these results (Sutton & Barto, 1999). This approach employs a feedback system to reward good behaviour and penalise bad behaviour. A self-driving car is a great example of reinforcement learning (Sutton & Barto, 2018).

#### *2.2.2.1 Support Vector Machine*

Support Vector Machine (SVM) is a commonly used machine learning algorithm that is renowned for its effectiveness in classification and regression tasks. It was first introduced by Cortes et al. (1995). It works by transforming the input data into a higher-dimensional space and finding the optimal hyperplane that maximises the margin between the classes.

A portion of the data called the support vectors are used to train the SVM algorithm. To determine the decision boundary, the data points that are closest to the hyperplane are considered. Convex optimization techniques, such as gradient descent, quadratic

programming, or interior point approaches, can be used to identify the best hyperplane for SVM, which is a convex problem. In comparison to other machine learning methods, it has several advantages, including the capacity to handle high-dimensional data, operate with non-linear decision boundaries, and generalise well to fresh data. Additionally, SVM is a strong algorithm that is less impacted by data outliers. Figure 2.1 illustrates that SVM is finding the optimal hyperplane which is H2.

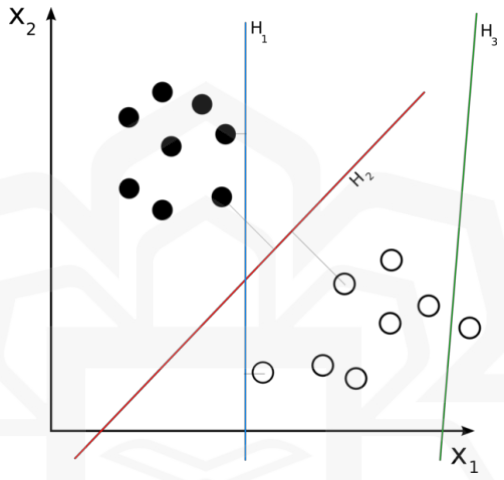



Figure 2.1 SVM finding hyperplane with maximum margin.

### *2.2.2.2 Logistic Regression*

Logistic regression is a statistical technique used to establish the connection between one or more independent variables and a categorical dependent variable. The goal is to model the probability of the dependent variable taking a particular value, given the values of the independent variables, using a logistic function to transform a linear combination of the independent variables into a probability value between 0 and 1.

When a dependent variable is categorical and one or more independent variables are also present, the relationship between the two can be modelled statistically using logistic regression. The objective is to use a logistic function to convert a linear combination of the independent variables into a probability value between 0 and 1 so that

the probability of the dependent variable having a specific value can be modelled given the values of the independent variables.

In order to forecast probabilities ranging from 0 to 1, this model converts a linear equation into a nonlinear sigmoidal curve. The formula for the equation is:

$$p = \frac{1}{1 + e^{-z}} \tag{2.1}$$

Where $p$ is the likelihood that the dependent variable would have the value 1, $z$ denotes a linear combination of the predictor variables, and $e$ denotes the natural logarithm's base. Logistic regression is often used in binary or multi-class classification problems and can also be used in survival analysis and other applications where the dependent variable is a probability. This model is computationally very fast.

### 2.2.3 Deep Learning

Bengio et al. (2015) defined deep learning as a subset of machine learning that relies on artificial neural networks (ANN) with representation learning. ANN is composed of interconnected nodes, or neurons, which work together to accomplish a specific objective. It is a type of machine learning model that draws inspiration from the structure and operation of biological neural networks in the human brain. Simply put, an ANN is a network of neurons as described by Schmidhuber (2015).

Each neuron in an ANN receives input from one or more neurons and produces output that is sent to the other neurons. The connections between neurons are weighted, which means that the strength of the connection determines how much one neuron's output influences the output of another. An ANN's connection weights between neurons are altered during training so that the network may learn to produce the correct output for a given input. Typically, an optimisation algorithm like gradient descent is used to accomplish this, which modifies the weights based on the difference between the expected

and actual results. To reduce the difference between the predicted and actual output, the weights are iteratively updated.

Neural networks come in different varieties, including recurrent neural networks (RNN), deep neural networks (DNN) and feedforward neural networks (FNN). A deep learning model is composed of multiple layers of neural networks (NN). Figure 2.2 depicts a fully connected ANN, which is made up of several layers of neural networks: an input layer, one or more hidden layers, and an output layer. This is often referred to as a dense neural network or dense layer in Keras (i.e., a popular deep learning library).

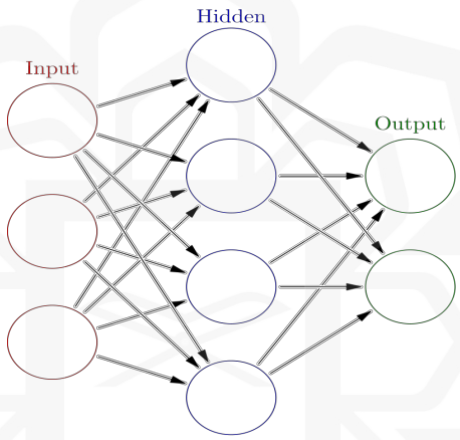



Figure 2.2 Dense Neural Network

### 2.2.4 Transfer Learning

Transfer learning is a strategy in ML that involves using a pre-existing model to solve a related task. This process includes adjusting certain parameters like the number of epochs, loss function, and learning rate to refine the model for the current task. The advantage of transfer learning is that it reduces the amount of data and computational power required to train a model for the specific task. Using a pre-trained model as a foundation enables the model to learn patterns and characteristics more efficiently and precisely than starting the training process from scratch. This is especially useful in cases where the target dataset is small or there are limited computing resources available.

Transfer learning is commonly used in NLP, computer vision (CV), and other machine learning applications where large datasets are available for pre-training models. Some examples of pre-trained models are resNet (He et al., 2016), BERT (Devlin et al., 2018), VGG (Simonyan et al., 2014). These models are trained with huge amounts of data and can be re-used for the same domain related task using transfer learning methods. For instance, resNet was trained with millions of images from imageNet database and it can be re-used for any image processing related task. Figure 2.3 illustrates the framework based on transfer learning for pneumonia detection.

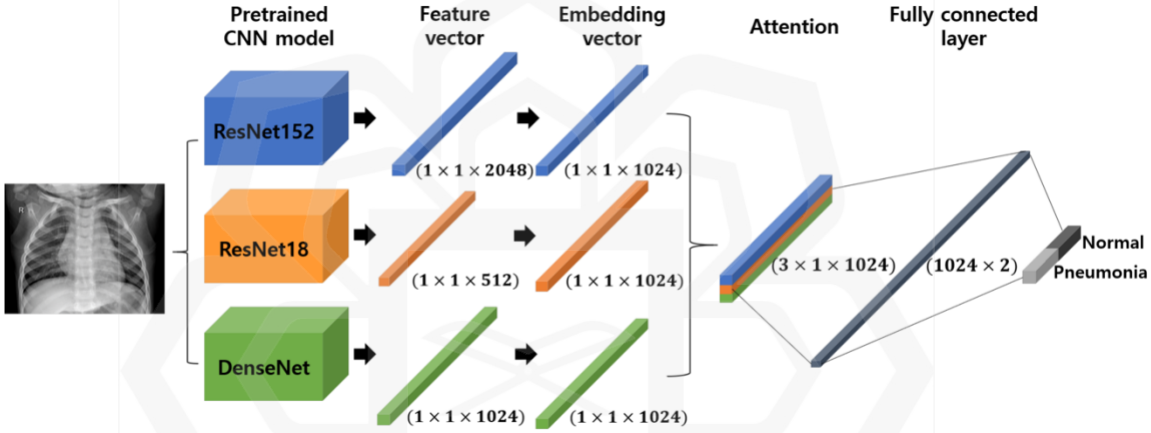



Figure 2.3 Transfer learning based framework for pneumonia detection. (Cha et al., 2021).

### *2.2.4.1 BERT*

BERT stands for Bidirectional Encoder Representation from Transformer. It is the most up-to-date embedding model published by Google (Devlin et al., 2018). BERT was trained on BooksCorpus which contains 800 millions words (Zhu et al., 2015) and English Wikipedia (2,500 millions words) datasets. This model can be re-used for sentence categorization, text generation, sentiment analysis and question answering.

Vaswani et al. (2017) introduced BERT which is a transformer-based model specifically designed to process sequential data. BERT is capable of performing a variety

of tasks, including translation and text summarization, by effectively handling the inherent complexity and variability of natural language data. It consists of encoder-decoder architecture. The output from the final encoder layer is processed by the decoder layers. All the encoder and decoder blocks are identical to each other. Figure 2.4 illustrates an example of a transformer.

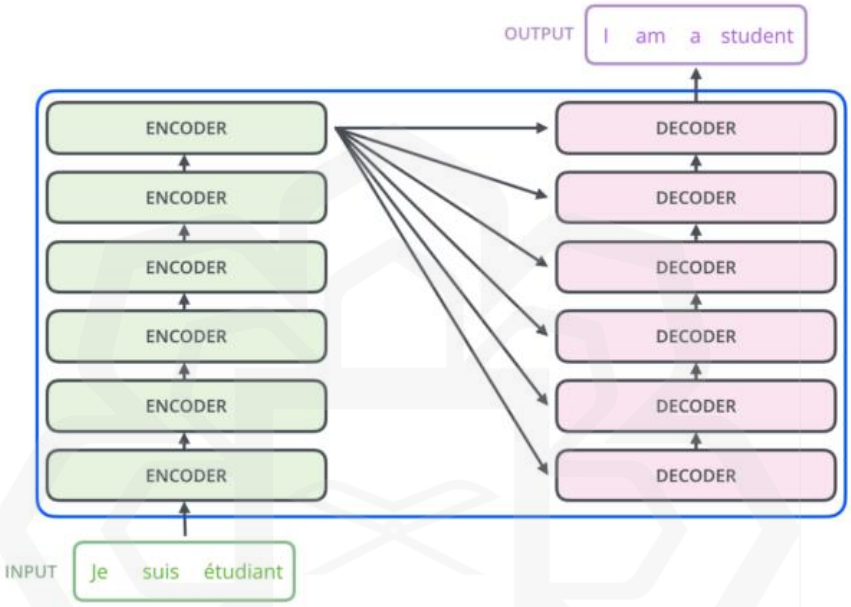



Figure 2.4 Transformer (Alammar, 2022)

There are two main components of an encoder. The first component is the self-attention layer followed by a Feed Forward Neural Network (FNN) layer. The input given in the encoder first flows through the self-attention layer. Self-attention layer assists the encoder to examine all the other input words before encoding a specific word. For example, to produce output vector Z of *x1* for input *X*, which consists of *x1*, *x2*, *x3*, the self-attention layer will look into all the other input words *x2* and *x3*. Finally, FNN will produce output for *Z* and pass it to the next encoder. Figure 2.5 depicts the encoder architecture.

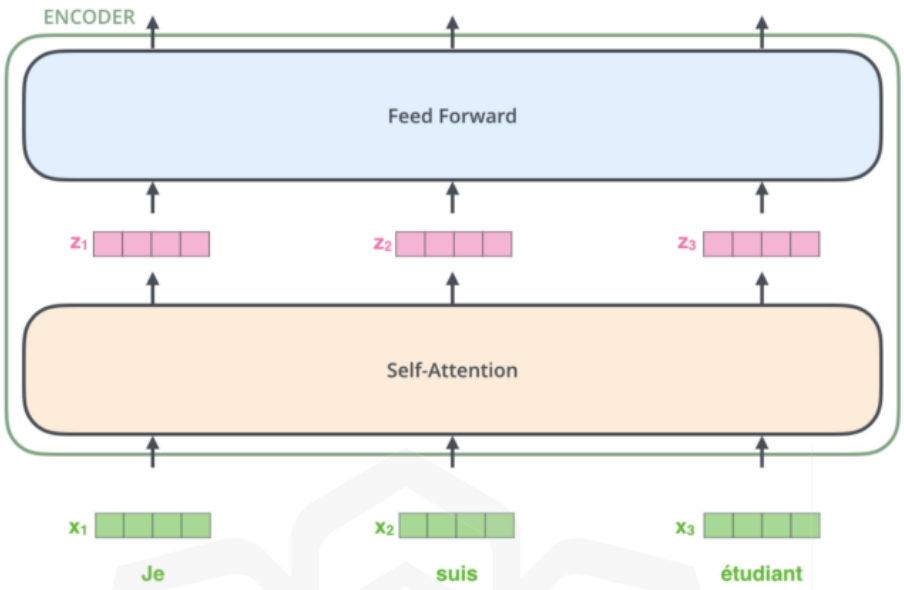



Figure 2.5 Encoder (Alammar, 2022)

BERT-small, BERT-base, and BERT-large are three distinct pre-trained BERT models that differ in terms of their size and intricacy. BERT-small stands as the most compact and rapid BERT model, boasting 6 encoder layers, 768 hidden units, 12 attention heads, and 125 million parameters. It serves as an optimal choice for applications prioritizing speed over top-notch performance, particularly in real-time scenarios. BERT-base represents the middle-ground BERT model, equipped with 12 encoder layers, 768 hidden units, 12 attention heads, and 110 million parameters. It offers a well-balanced solution for most tasks, striking a harmony between performance and speed. BERT-large takes the crown as the most extensive and intricate BERT model, featuring 24 encoder layers, 1024 hidden units, 16 attention heads, and 340 million parameters. It excels in performance across a wide range of tasks, but at the expense of being the slowest and most computationally demanding option.

Before using BERT as transfer learning, some of the hyperparameters can be modified as needed. Hyperparameters in deep learning are essentially settings that govern the training process and determine the specific values that the learning algorithm will ultimately assign to a model's parameters. These hyperparameters are established prior to the commencement of training, and they play a pivotal role in instructing the learning

algorithm on how to adapt the model's parameters throughout the training process. While the training process unfolds, the parameters are continuously adjusted, and the end result of this iterative process forms the foundation of the final model. In deep learning, several examples of these hyperparameters are as follows:

*Learning rate*: The learning rate serves as a crucial factor in dictating the pace at which the model acquires knowledge. A higher learning rate can accelerate convergence, but it also carries the risk of introducing instability and overfitting.

*Batch size*: The batch size is responsible for determining the number of data samples utilized to update the model's parameters during each training iteration. A larger batch size can enhance computational efficiency, although it may render the model more sensitive to noise present in the data. Typically batch size is set to 32.

*Number of epochs*: The number of epochs influences how frequently the entire training dataset is processed by the model during training. A larger number of epochs can potentially lead to improved model performance, but it comes at the cost of extended training time.

*Optimizer*: The optimizer governs the mechanism through which the model's parameters are adjusted during training. A plethora of optimizer options is available, each possessing distinct strengths and weaknesses. One of the optimizer is known as *AdamW*. It is an optimization technique that utilizes adaptive estimates of both first-order and second-order moments while incorporating a weight decay mechanism following the principles outlined in the 2017 paper titled 'Decoupled Weight Decay Regularization' (Loshchilov et al., 2017). It represents an enhancement to the Adam optimizer, specifically designed to rectify an issue in *Adam's* handling of weight decay. In the original *Adam* optimizer, weight decay is applied to the gradients prior to parameter updates, potentially causing convergence issues, especially when dealing with extensive models. Hence *adamW* is the recommended optimizer for BERT.

*Regularization parameters*: Regularization parameters are deployed to prevent overfitting, which occurs when a model becomes overly attuned to the training data, impairing its ability to generalize to new data. Common regularization techniques encompass L1 and L2 regularization. Another popular regularization technique in deep learning is called *Early stopping*. It involves monitoring a model's performance on a validation dataset during training and stopping when the performance worsens, typically due to an increase in validation error. This helps prevent overfitting, saves training time, and improves generalization. Key steps include splitting the data, defining a stopping criterion, and saving the best-performing model for testing. Properly tuning hyperparameters is crucial for effective *early stopping*.

*Dropout*: Dropout stands as a technique that randomly deactivates certain neurons during training. This technique helps combat overfitting by compelling the model to learn more resilient features.

*Loss Function*: A loss function is a mathematical measure of how well a machine learning model's predictions match the actual target values. It guides the model's training by quantifying the error to minimize, helping the model improve its performance. The type of loss function used depends on the specific task, such as regression, classification, or sequence-to-sequence modeling. One of the popular loss functions is the *binary cross-entropy*. It is also known as log loss, is a key loss function in machine learning and deep learning for binary classification tasks. It quantifies the disparity between predicted binary outcomes and true binary labels. This loss encourages models to predict probabilities that align with the actual labels, penalizing incorrect predictions with high magnitudes. The formula for *binary cross-entropy* loss is used to calculate this discrepancy and is applied to each example in the dataset. The goal during training is to minimize this loss, often through optimization methods like *gradient descent*, to improve model accuracy in binary classification problems.

### 2.2.5 Natural Language Processing

Natural Language Processing (NLP) is a branch of machine learning that gives computers the ability to comprehend, examine, alter, and even create human language. Liddy (2001) defines NLP as: "a theoretically motivated range of computational techniques for analysing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications."

The intricacy and ambiguity of human language make NLP a difficult field to work with. Combining methods from computer science, linguistics, and machine learning is necessary. NLTK, spaCy, and CoreNLP are a few of the well-known NLP frameworks. NLP is frequently used in the following contexts:

- Text classification: grouping text into different topics or themes.
- Named entity recognition: identifying and classifying named entities (such as people, places, and organisations) in text.
- Machine translation: automatically translating text from one language to another.
- Speech recognition: transcribing spoken language into text.
- Question answering: robotically responding to questions.
- Sentiment analysis: A technique of computationally recognising and classifying the attitudes represented in a text, such as those that are positive, negative, or neutral.

#### *2.2.5.1 Sentiment Analysis*

Sentiment analysis is a branch of linguistics that includes computational linguistics, text mining, and natural language processing. It makes use of data mining, machine learning and computational linguistics technologies to study subjective aspects of text, such as views and emotions. Whether views are expressed explicitly or implicitly, a sentiment might have a positive, negative, or mixed polarity (Mejova, 2009). In recent years, the popularity of

this technique has grown significantly due to the emergence of social media and the need to analyse user-generated content.

Numerous studies have been done on sentiment analysis, examining different approaches and uses. Pang and Lee (2008) compared the performance of various machine learning algorithms, such as maximum entropy, support vector machines, and naive bayes, and discovered that these methods performed well in sentiment analysis tasks, especially when combined with feature selection methods. Thelwall and Buckley (2013) looked into the use of sentiment analysis to social media data, particularly Twitter. They found that using Twitter sentiment, sentiment analysis was able to predict election results with accuracy, highlighting the usefulness of this method for political forecasting. Alamoodi et al. (2021) conducted a systematic review that investigates the several ways that sentiment analysis is used to address vaccine reluctance. The review examines how sentiment analysis can be applied to address vaccination reluctance using a multi-perspective approach and an analysis of pertinent literature. According to the study, sentiment analysis can be used to understand the causes of vaccine reluctance and create ways to combat it. It also points out several difficulties and restrictions that must be overcome in order to guarantee sentiment analysis's usefulness in this context.

#### *2.2.5.2 TF-IDF*

Term Frequency-Inverse Document Frequency or TF-IDF is a technique used in information retrieval and NLP to assess the significance of a term in a document. It is frequently used in text classification, text clustering, sentiment analysis and search engines. This feature is often used in machine learning models for text classification and sentiment analysis (Fransiska et al., 2020; Mee et al., 2021). A term's TF-IDF value is determined by two variables:

- Term Frequency (TF): The frequency with which a term appears in a document is measured by its term frequency (TF). The calculation is made by dividing the number of terms in the document by the frequency of the phrase.

- Inverse Document Frequency (IDF): An indicator of a term's significance to the entire collection of documents is the inverse document frequency (IDF). The IDF of a term is determined by taking the logarithm of the total number of documents in a corpus, and then dividing it by the number of documents that contain the term. Essentially, this means that the IDF of a term increases when it appears in fewer documents, and decreases when it appears in more documents.

To obtain the TF-IDF value of a term, its TF and IDF values are multiplied together.

## 2.3 VACCINE HESITANCY DETECTION TECHNIQUES

### 2.3.1 Analytical Techniques

This section compiles the analytical techniques employed for the identification of vaccine hesitancy within the realm of social media. It encompasses the strategies employed to meticulously examine data and derive substantive insights. These methodologies encompass conventional statistical approaches, data visualization techniques, and exploratory data analysis.

The effect of social media on vaccine reluctance during the COVID-19 pandemic is discussed, and a real-time big data analytics approach is suggested to examine social media sentiment on vaccination by Bari et al. (2022). The framework tracks sentiment and content themes of tweets in real-time using sentiment analysis. According to the findings, there is an association between sentiment score and vaccination rates with a one week lag. According to the study, social media monitoring of vaccine sentiment could help medical practitioners get ready for vaccination campaigns and talks. Analysing the interactions

between particular communities, vaccine attitudes, and vaccination rates will require more research.

Vaccine-hesitant groups and their geolocations were analysed by Ruiz et al. (2021). Data from twitter was used to determine anti-vaccine communities and their geolocation. They found 420 significant Twitter users after collecting 139,433 messages. To identify Twitter discussion patterns, a semantic network technique was used. This method shows that identifying dominant social media users could help detect vaccine hesitant communities. They also used IBM Watson NLP to conduct sentiment analysis on vaccines to determine the percentage of positive, neutral, and negative sentiment.

A monitoring tool known as Vaccinpraat for COVID-19 vaccination in the Dutch language was developed by Lemmens et al. (2021). The tool provides statistical insights and argumentation detection for anti-vaccine posts shared on Facebook and Twitter. They used named entity identification, fine-grained emotion analysis, and author profiling techniques to process the content of vaccine-related messages. Argument detection experiments revealed that this is more difficult than stance detection, and more research is required in this area.

An analysis of COVID-19 vaccine sentiments and opinions on Twitter was done by Yousefinaghani et al. (2021). A total of 4,552,652 tweets were collected from January 2020 to January 2021. To perform the sentiment analysis, the rule-based analysis tool, VADER, was used. The sentiment analysis results showed that 34% of the tweets were positive (34%) towards vaccination, while vaccine hesitancies and objections were more dominant than vaccine interests. Further analysis of the dataset was done by categorising randomly selected 500K tweets into anti-vaccine, hesitant, and pro-vaccine which showed that the number of tweets for vaccine categories differed from region to region. Almost identical research was done by Na. et al. (2021). In this research, the authors collected tweets from the UK and USA related to the COVID-19 vaccine. The sentiment analysis was done by using the VADER tool. The results showed that online celebrities can influence public opinion on vaccines. Almost 40% of the residents from both countries have negative

outlooks towards the COVID-19 vaccines. This study also found that among all the COVID-19 vaccines, the Pfizer vaccine has the most positive public sentiment. A topic modelling using the Latent Dirichlet Allocation (LDA) model was performed at the end of the research and found that these nations are open to discussing their perspectives and concerns about vaccination.

Analysis of the anti-vaccination movement from social media posts was performed by Garay et al. (2019). The K-means clustering algorithm and the Valence Aware Dictionary and sEntiment Reasoner (VADER) sentiment analyser is used to perform the analysis. Data was collected from public anti-vaccination group's posts on Twitter and Facebook. The Term Frequency-Inversed Document Frequency (TF-IDF) vectorization method is used as a feature for K-means clustering with 10 clusters. A Silhouette score is used to evaluate the cluster's performance. The overall silhouette score was 0.013540022. A VADER sentiment analyser is used to further describe each cluster. The silhouette score indicates that the points are close to the decision boundaries.

Table 2.1 provides the summary of five studies on sentiment analysis on social media on vaccine hesitancy as discussed earlier. These research used a variety of vaccine hesitancy analysis approaches, including sentiment analysis, named entity identification, fine-grained emotion analysis, author profiling techniques, K-means cluster, VADER sentiment analyzer, and LDA. The sources of the datasets used in these studies include Facebook and Twitter.

Table 2.1 Analytical techniques for the detection of vaccine hesitancy

| No | Research Work | Detection technique | Source of Dataset |
|---|---|---|---|
| 1 | Bari et al., 2022 | Sentiment Analysis | Twitter |

| | | | |
|---|---|---|---|
| 2 | Ruiz et al., 2021 | Semantic network approach, Sentiment analysis using IBM Watson NLP | Twitter |
| 3 | Lemmens et al., 2021 | Named entity identification, fine-grained emotion analysis and author profiling techniques | Facebook, Twitter |
| 4 | Yousefinaghani et al., 2021 | VADER sentiment analyzer, and LDA | Twitter |
| 5 | Garay et al., 2019 | K-means Cluster and VADER sentiment analyzer | Facebook |

### 2.3.2 Machine Learning Techniques

This section delves into the utilization of machine learning techniques, wherein the authors conducted training and testing of machine learning algorithms dedicated to the identification of vaccine hesitancy within the context of social media.

Qorib et al. (2023) utilised social media data gathered from publicly streamed live tweets to analyse COVID-19 vaccine reluctance. The data were analysed using three sentiment calculation approaches: Azure Machine Learning, VADER, and TextBlob. Five machine learning models were trained using various vectorization techniques and vocabulary normalisation techniques. The study found that lemmatization and potter stemming together improved model performance. Linear Support Vector Classification (LinearSVC) model with TF-IDF vectorization technique outperformed all the other models. The accuracy, precision, recall, and F1 scores of this model were 0.96752, 0.96921,

0.92807, and 0.94702 respectively. Additionally, the study discovered that COVID-19 vaccine hesitancy gradually decreases over time.

Research work on the detection of pro-vaccine and anti-vaccine behaviours from social media using machine learning was conducted by Argyris et al. (2021) and Yuan & Crooks (2018). The dataset was collected from Twitter and categorised as pro-vaccine, anti-vaccine and neutral. Argyris et al. (2021) trained a Logistic Regression (LR) classifier that had an F1-score of 95.7. Yuan and Crooks (2018) used a SVM to investigate the MMR (Measles, Mumps, and Rubella) vaccine outlook. The best performance on the test dataset was an F1-score of 73.13.

A study on applying machine learning to analyse anti-vaccination on Tweets was performed by Taeb et al., (2021). Sentiment analysis technique was used to evaluate the COVID-19 anti-vaccine related tweets. The performance of several models using various configurations and training datasets were investigated in this study. Automatic topic modelling was done with LDA and BERT, and sentiment analysis was done with TF-IDF and logistic regression.

An experiment on the detection of reliable or not reliable web pages about early childhood vaccination was conducted by Meppelink et al. (2021). In their research, they used a supervised machine learning approach. First, they classified the webpages manually into two categories: reliable or not reliable. Reliable means the information on the web page related to early childhood vaccination is positive, while not reliable means the information is not correct or negative. A Naïve Bayes (NB) classifier was trained and tested. The test F1-score of the model was 88.0.

Social media sentiments on disease and vaccination in the Spanish language were studied by a project called MAVIS (Rodríguez-González et al., 2020). They collected data from Twitter and Instagram and used the Synthetic Minority Oversampling Technique (SMOTE) method to balance the dataset. After training, the Receiver Operating Characteristic (ROC) score of a Random Forest (RF) classifier was 88.0. An almost

identical experiment was done by Du et al. (2017). They conducted the experiment using Human Papillomavirus (HPV) related tweets. The dataset was manually classified and trained with SVM. Using the optimised features set, the model achieved an F1-score of 74.42. Another identical study to detect vaccine hesitancy in social media was done by Piedrahita-Valdés et al. (2021). They collected 1,499,227 tweets within 8 years. An SVM model was trained and tested. The accuracy for the test data was 85% in terms of detecting positive and negative sentiments. A thorough examination of the dataset revealed that both positive and negative sentiments towards vaccination increased over time, while neutral sentiment decreased.

Table 2.2 provides a detailed summary of six different research studies that used machine learning techniques to identify instances of vaccine hesitancy outlook in social media. Machine learning methods like LinearSVC, LR, SVM, and NB were employed as the detecting strategies. Twitter served as the dataset's primary data sources. Mainly, F1-scores were used to evaluate the effectiveness of the detection strategies, with F1-scores ranging from 0.73 to 0.94.

Table 2.2 Machine learning techniques for the detection of vaccine hesitancy

| No | Research Work | Detection Technique | Source of Dataset | Performance |
|---|---|---|---|---|
| 1 | Qorib et al., 2023 | LinearSVC | Twitter | F1-score: 0.94 |
| 2 | Argyris et al., 2021 | Logistic regression | Twitter | F1-score: 86.9 |
| 3 | Taeb et al., 2021 | Logistic regression | Twitter | N/A |

| | | | | |
|---|---|---|---|---|
| 4 | Meppelink et al., 2021 | Naïve Bayes | Web pages | F1-score: 91.0 |
| 5 | Rodríguez-González et al., 2020 | Support vector machine | Twitter | ROC: 88.0 |
| 6 | Yuan & Crooks, 2018 | Support vector machine | Twitter | F1-score: 73.0 |

### 2.3.3 Deep and Ensemble Learning Techniques

This section presents a comprehensive literature review of vaccine hesitancy detection methods, with a particular focus on the deep and ensemble learning techniques. Deep learning techniques encompass the utilization of conventional deep learning algorithms, such as Neural Networks (NN) and Long Short-Term Memory (LSTM), whereas ensemble learning techniques involve the integration of two or more machine learning or deep learning models into the detection process.

The application of multimodal deep learning for the identification of medical falsehood about vaccination was proposed by Wang et al. (2021). The authors used Instagram to collect the data for this study, which included photographs, hashtags, and textual contents. In this study, a deep learning network that combines visual and textual data was developed. To assist the model in focusing on the crucial parts of a post that signify anti-vaccine messaging, a new semantic and task-level attention mechanism was constructed. This suggested model can generate accurate integrated characteristics for predictions. In addition, an ensemble strategy is adopted to boost the final forecast accuracy even more. The results from 30 experiments showed that the final model beats other similar models in terms of testing accuracy, suggesting that it can identify a large number of anti-vaccine statements posted on a daily basis.

The study on the use of Twitter to analyse public opinions towards vaccination was done by Baru et al. (2019). Initially, the tweets were divided into relevant and irrelevant. The relevant tweets were then split into negative, positive, and neutral. A neural network model was built and trained. The best accuracy was an F1-score of 58.0.

Stigmatised behaviour towards vaccination using predictive modelling was proposed by Yoo et al. (2019). The data was collected from Facebook comments that were separated into two groups: negative and positive. The dataset was used to train and evaluate a few machine learning models. The model which was built using fastText outperformed all the other models.

Vaccination behaviour detection entails predicting whether a person has gotten or will receive a vaccine. Joshi et al. (2019) constructed an ensemble learning model using LR, SVM, and RF algorithms. The data was gathered from tweets on Twitter. At least two classifiers must classify a tweet as positive for it to be considered positive. This method's F1-score was reported to be 80.75. Furthermore, to identify vaccination behaviours, few deep learning models such as Convolutional Neural Network (CNN), dense neural network (DNN) and Long Short-term Memory Language Model (LSTM-LM) were trained and tested separately. With an F1-score of 80.87, LSTM-LM outperformed the other deep learning models.

Table 2.3 summarises four different research works that were discussed earlier for vaccine hesitancy detection using deep and ensemble learning. Each work used a different detection technique and methodology, and the source of dataset varied among Twitter, Facebook, and Instagram. The performance metric also varied among the works, with F1-score and Accuracy being used to measure performance. The highest performance was achieved by Wang et al. in 2021, with an accuracy of 97% using an ensemble learning technique with NN and SVM.

Table 2.3 Deep and Ensemble learning techniques for vaccine hesitancy.

| No | Research Work | Detection Technique | Source of Dataset | Performance |
|---|---|---|---|---|
| 1 | Wang et al., 2021 | NN, SVM | Instagram | Accuracy: 97.0 |
| 2 | Baru et al., 2019 | NN | Twitter | F1-score: 58.0 |
| 3 | Yoo et al., 2019 | FastText | Facebook | Accuracy: 75.0 |
| 4 | Joshi et al., 2019 | DNN, BiLSTM, CNN, LSTM-LM | Twitter | F1-score: 80.87 |

### 2.3.4 Transfer Learning Techniques

This section presents a comprehensive literature review pertaining to vaccine hesitancy, with a particular focus on transfer learning techniques, notably the BERT model.

A study conducted by Marcec et al. (2022) explored how the spread of COVID-19 has been limited by vaccines, but some people are hesitant to get vaccinated due to misinformation. The study analysed public sentiments and hesitancy towards COVID-19 vaccines on Instagram posts. Over 10k comments and captions were retrieved from vaccine hashtags, translated into English, and assigned a polarity score (positive, negative, neutral). The study found that the posts related to hashtag Covaxin had 71.4% positive, Covishield had 64.2% positive, and Sputnik had 55.8% positive sentiments. Furthermore, BERT model was fine-tuned and used for vaccine sentiment prediction. The best F1-score obtained by this model was 84.60. The findings suggest that understanding vaccination perceptions

through social media could be useful for public health officials to promote positive marketing and reduce negative marketing. The paper also suggests some future directions for research on this topic.

Vaccine hesitancy and argumentation detection on Dutch language using BERT was studied by Lemmens et al. (2021). The model was developed using 2.8 million Dutch tweets related to COVID-19 that were posted in 2021, and it was constructed by expanding the pre-training phase of RobBERT, as described in a study by Delobelle (2020). The two models were tested on two tasks: (1) binary vaccine reluctance detection and (2) detection of justifications for vaccine hesitation, in order to compare their performances. To demonstrate cross-genre proficiency in both tasks, data from both Twitter and Facebook was used. The best accuracy for the model was an F1-score of 73.0.

Liu et al. (2021) conducted sentiment analysis using Twitter dataset. BERT, BiLSTM, SVM and NB models were trained and evaluated. The evaluation matrices showed that BERT outperformed all the other models while detecting vaccine sentiment. Similar study was done by To et al. (2021). This study also found that BERT performed better in terms of vaccine hesitancy detection compared to other techniques. An F1-score of 95.5 was achieved by the BERT model.

A study on COVID-19 vaccine hesitancy in the month following the start of the vaccination process was conducted by Cotfas et al. (2021). BERT was adopted as one of the classification methods. The optimal parameters for the built NLP pipeline were determined using a grid search approach. The performance of the BERT model in terms of accuracy was 75% which was slightly better than the other classical machine learning techniques (SVM, RF).

Transfer learning was used to conduct sentiment analysis on Twitter data for the HPV vaccine by Zhang et al. (2020). They experimented sentiment analysis with a fine-tuned BERT model which scored the best among other methods in analysing vaccination

attitudes. The F1 score of 76.9 was achieved by the model. Also, a NN model was trained which didn't perform well as compared to fine-tuned BERT model.

Table 2.4 lists research works that have developed vaccine hesitancy detection models using transfer learning with BERT on social media datasets such as Instagram, Twitter, and Facebook. The performance of the models is evaluated using F1-score or accuracy. To et al. (2021) achieved the highest F1-score of 95.5 on the Twitter dataset, while Lemmens et al. (2021) having the lowest F1-score of 73.0 on both Twitter and Facebook datasets.

Table 2.4 Transfer learning techniques for vaccine hesitancy detection.

| No | Research work | Source of dataset | Performance |
|---|---|---|---|
| 1 | Marcec et al., 2022 | Instagram | F1-score: 84.60 |
| 2 | Liu et al., 2021 | Twitter | F1-score: 79.2 |
| 3 | To et al., 2021 | Twitter | F1-score: 95.5 |
| 4 | Cotfas et al., 2021 | Twitter | Accuracy: 75 |
| 5 | Lemmens et al., 2021 | Twitter, Facebook | F1-score: 73 |
| 6 | Zhang et al., 2020 | Twitter | F1-score: 76.9 |

### 2.3.5 Research Gap in Previous Research Studies

It is evident that most of the research studies considered the use of only a single social media platform (Zhang et al., 2020; Cotfas et al. 2021; Baru et al., 2019). Figure 2.6 also represents the tendency of using mono-platform data when building model for vaccine hesitancy detection (Hossain & Sulaiman, 2022).

Only one study attempted to use data from both Twitter and Facebook for vaccine hesitancy detection (Lemmens et al., 2021). However, there are a few limitations in their research. One of the limitations was that the research was conducted to detect vaccine hesitancy in Dutch language only. Another limitation was the size of the dataset. A total of 8,800 tweets and 5,200 Facebook posts were used in their research which is relatively small for model training and validation. The best accuracy obtained was an F1-score of 73. One interesting fact found in this study is that training the model using data from two different platforms (i.e., Facebook and Twitter) improves the model's overall performance. This is an indication that training the model using data from multiple social media platforms could further improve the model performance. The authors also concluded their study by recommending the use of more data from multiple social media sources for future studies. Most of the research was conducted by using twitter dataset only. There are also a lack of open source models and datasets which can be reused by other researchers.

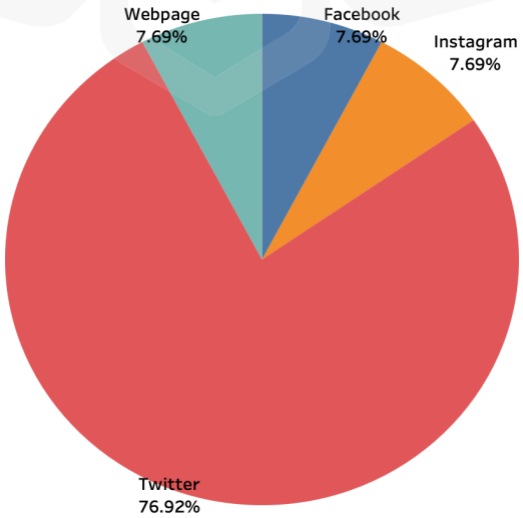



Figure 2.6 Overview of social media usage as a data source (Hossain & Sulaiman, 2022).

Recently, transfer learning methods using BERT for vaccine hesitancy detection have gained significant popularity among research communities (Zhang et al., 2020; Liu et al., 2021; To et al., 2021; Lemmens et al., 2021). A very recent systematic review by Hossain & Sulaiman, (2022) also found BERT to be the second most efficient method behind SVM for detecting public outlook towards vaccination. Figure 2.7 depicts the modelling techniques used in vaccine sentiment detection (Hossain & Sulaiman, 2022).

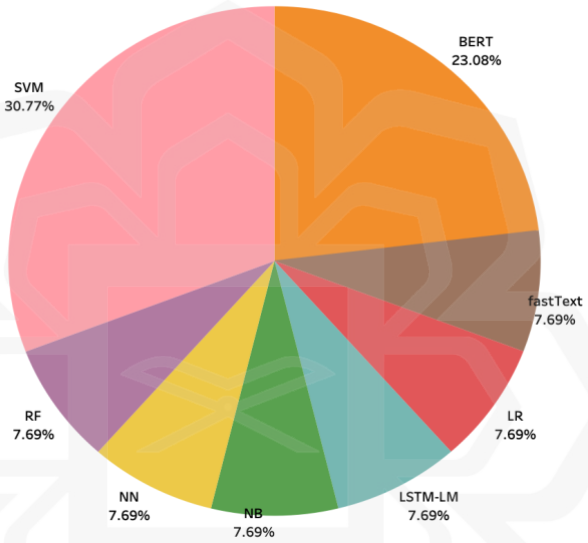



Figure 2.7 Modelling techniques used in vaccine sentiment detection (Hossain & Sulaiman, 2022).

Figure 2.8 shows the top models that perform well when detecting vaccine hesitancy from social media which include BERT, SVM and LR (Hossain & Sulaiman, 2022).

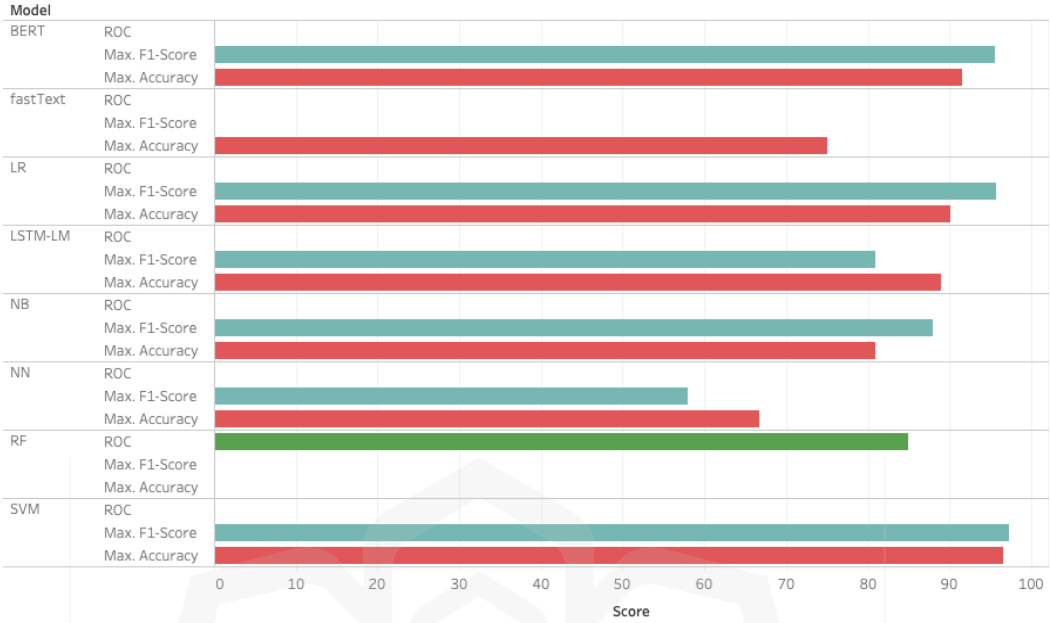



Figure 2.8 Top results by models (Hossain, & Sulaiman, 2022).

## 2.4 SUMMARY

This chapter serves as a foundational resource for understanding the theoretical underpinnings and various methods used in the critical task of identifying and addressing vaccine hesitancy, with a particular focus on transfer learning techniques like BERT. Vaccine hesitancy has emerged as a significant concern on major social media platforms, posing potential risks to public health. To address this challenge, a range of cutting-edge techniques has proven effective in detecting vaccine hesitancy. These include the utilization of advanced models like BERT, as well as traditional machine learning algorithms such as SVM and LR. However, a notable research gap exists in the integration of data from multiple social media platforms as explained in Section 2.3.5.

# CHAPTER THREE
# METHODOLOGY

## 3.1 INTRODUCTION

In Chapter 2, we have presented a comprehensive review on the theoretical background underlying the vaccine hesitancy detection techniques. This chapter provides an exposition of the experimental research design and the comprehensive workflow for conducting the experiment. The organisation of this chapter is as follows: Section 3.2 presents an overview of experimental research in the field of Computer Science. Section 3.3 delves into the Modelling framework, which entails data collection, data pre-processing, data analysis, and model selection and model evaluation, while section 3.4 provides a succinct summary of the entire experimental framework.

## 3.2 EXPERIMENTAL RESEARCH DESIGN

According to Creswell (2012), research is a systematic approach that employs various techniques to gather, analyse, and interpret data with the aim of obtaining a scientific understanding of a phenomenon or topic. Quantitative and qualitative research are two fundamental methodologies employed in research. The experimental research design, however, bridges the gap between these two approaches as it aims to determine whether a set of practices influences the outcome of measurements under specific experimental conditions (Creswell, 2012). Experimental research design is supposedly quantitative in nature. Deep learning, a subfield of machine learning, is centred around experimentation and relies on a trial-and-error approach to achieve the optimal correlation of parameters (i.e., weight and bias values) while minimising the cost function of gradient descent computations (Goodfellow, 2016 & Kelleher, 2019). Given the requirement of several experiments when using deep learning techniques, the experimental research design is deemed the most suitable methodology for this research.

## 3.3 PROPOSED APPROACH

Figure 3.1 illustrates the comprehensive workflow of vaccine hesitancy detection framework. The initial phase of this workflow entails the collection of data from diverse social media platforms, encompassing Facebook, Twitter, and Reddit. Subsequently, the data undergoes preprocessing and feature extraction. The subsequent step entails model selection, with this research exploring three distinct models: SVM, LR and BERT.

Following model selection, the models are employed to determine the sentiment associated with vaccine hesitancy. Lastly, an evaluation of the models is conducted, employing various metrics including accuracy, F1-score, recall, and precision. This comprehensive workflow enables the systematic detection and assessment of vaccine hesitancy sentiment within social media discourse.

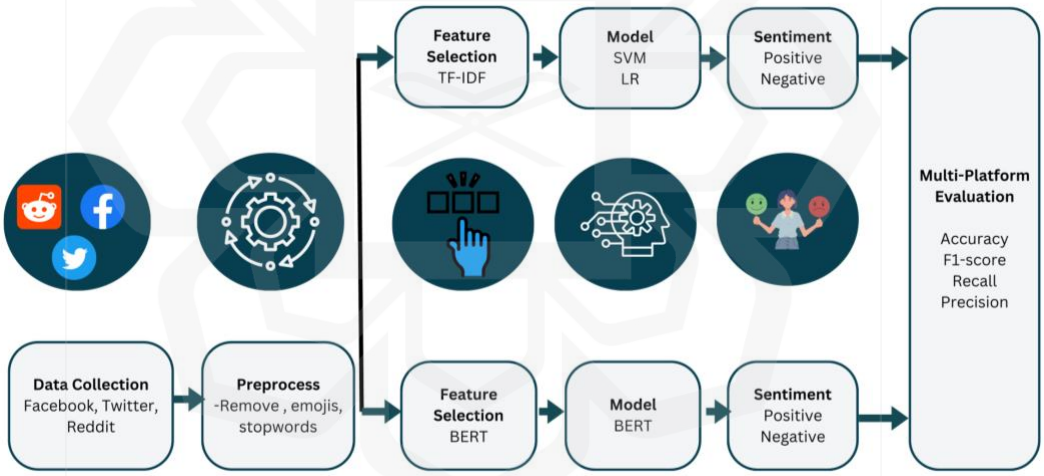



Figure 3.1 Overall workflow of vaccine hesitancy detection approach.

### 3.3.1 Data Collection

The first step in this research study is data collection. Several reliable sources have been identified, including anti-vaccine datasets from Facebook, Twitter, and Reddit. For this research, we exclusively collected labelled datasets related to vaccination, as we used a supervised machine learning modelling approach for vaccine hesitancy detection.

The Facebook anti-vaccine dataset, published by Kaggle (Helyar, 2018), comprises 89,867 posts related to anti-vaccine with 154 columns. The most significant columns include text and anti-vax, with the latter being labelled as true or false.

A Twitter anti-vax dataset published by Hayawi et al. (2022) were also collected, which consists of 15 million tweets related to anti-vaccine and vaccine misinformation. The data was gathered during the period spanning from November 2020 to July 2021. Out of these, only 15k tweets were manually annotated by medical experts. However, the texts of the tweets were missing, and only the tweet IDs were provided. To address this, we rehydrated this 15k annotated dataset with the Twitter IDs provided. Rehydration refers to the process of collecting full tweets through Twitter IDs by sending them to the Twitter live site.

Additionally, we obtained more Twitter vaccine hesitancy-related data from Kaggle, which were published by Fullmoondatascience (2021) and Bobhe (2022), respectively. Bobhe's (2022) data collection spanned from April 1st to April 7th, 2021. Both of these datasets were annotated as anti-vax or neutral. Out of all these datasets, only annotated data were kept and further preprocessed later. The total number of annotated data obtained from Twitter sources were 15k by Hayawi et al. (2022), 91k from Bobhe (2022) and 6k by Fullmoondatascience (2021).

Furthermore, we also gathered the Reddit anti-vaccine and lockdown sceptic dataset published by a Kaggle user named Lexyr (2020). This dataset, obtained from SocialGrep, contains 10 columns with over 2 million comments, annotated as -1 to 1, representing the most negative to positive outlook towards vaccination. Only labelled data were retained and pre-processed according to the preprocessing techniques mentioned in Section 3.3.3. Following the preprocessing, the total number of data were reported and analysed in Section 4.3 in detail.

### 3.3.2 Data Preprocessing

In every machine learning related research, data preprocessing is a key step. Given that our dataset is a collection of data from multiple sources, it is important to process the data accurately. Our proposed model requires only two main features: vaccine-related social media comments and their labels, such as positive or negative outlook. To this end, in the first step of data processing, we dropped columns that are not related to these two features. Subsequently, further processing of the data was carried out as outlined below.

1. *Removing Punctuation and Stopwords:* Punctuations such as double quote, comma, full stop were removed from the dataset. Stopwords were also removed. Stop words are commonly used in English such as “the”, “and”, “is”. Removing punctuation and stopwords make data less noisy and easy to analyse.

2. *Removing emojis and URLs:* Emojis and urls are not useful features for machine learning modelling. Hence, they have been removed.

3. *Dropping Null and duplicated rows:* The rows with missing values were removed. For rows with duplicate text values, only one row is kept. Null and duplicate values have adverse effects on machine learning modelling, specially when data is splitted into train and test sets.

4. *Normalization*: Finally, the whole dataset was converted into lowercase. Lowercase is part of data normalisation and it makes the text data analysis more efficient.

### 3.3.3 Data Analysis and Visualization

Data analysis and visualisation are two critical components of any data science or machine learning project. Data analysis involves using various statistical and mathematical techniques to explore, summarise, and draw insights from a given dataset. Visualisation,

on the other hand, involves creating graphical representations of data to facilitate understanding and communication of insights.

There are various tools available for data visualisation in Python which includes: Matplotlib, Wordcloud and Plotly. In this research, data analysis and visualisation were used to understand the distribution of data by platform, the distribution of data by sentiment and the top 50 words used in Facebook, Twitter and Reddit. Wordcloud visualisation was utilized for each social media platform by their sentiment labels. The result of the analysis is discussed in the following chapter.

### 3.3.4 Model Selection

In this research, the transfer learning method using BERT was applied to detect vaccine hesitancy posts from three (3) different social media platforms. BERT is a transformer-based machine learning model. Transformers utilise attention mechanisms to convert one sequence to another while eliminating the need for recurrence. This is accomplished by replacing the recurrences with attention. Due to the elimination of sequential dependency on previous words, this architecture allows a model to be trained more efficiently. This model also employs bidirectional transformer training on large amounts of computational resources and data. By reading the entire sentence simultaneously, the transformer encoder enables the model to grasp the context of a word through its neighbouring words, resulting in faster performance. Since the BERT model has already been pre-trained with a huge dataset, only small adjustments to the hyperparameters of the BERT model need to be made in the vaccine hesitancy detection tasks and its last few layers to produce the final output.

The sentiment analysis process by BERT is illustrated in Figure 3.2. This process involves several crucial steps, including text input, BERT model processing, and classification layer analysis. In the first step, the text input is processed by the input layer, which prepares the text for analysis by the BERT model. The BERT model itself then extracts BERT features from the input text. These features are designed to capture

important contextual information about the text, including the relationships between words and phrases.

Following the BERT feature extraction, the classification layer employs a softmax function to classify the text as having a positive or negative sentiment. This step involves analysing the BERT features extracted by the model to determine the overall sentiment of the text. Overall, the sentiment analysis process by BERT involves a series of steps aimed at accurately analysing the sentiment of text inputs.

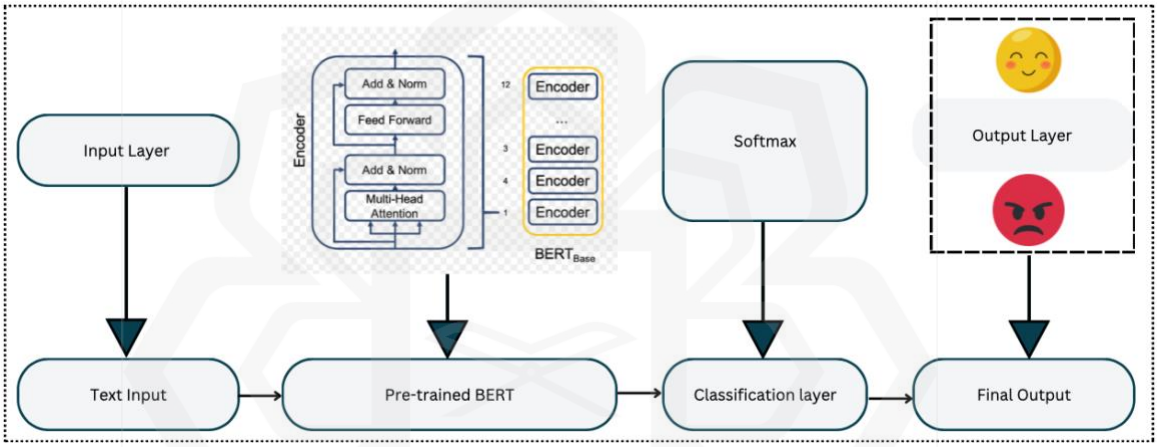



Figure 3.2 Sentiment Analysis with BERT

Furthermore, to compare the performance of the BERT model for vaccine hesitancy detection, SVM and LR models were also employed. The sentiment analysis workflow for these models is similar to that of the BERT model, with the exception of the feature input. In place of BERT features, we used a TF-IDF vectorizer for SVM and LR. By utilising SVM and LR models in combination with BERT, we can evaluate the effectiveness of multiple sentiment analysis techniques for vaccine hesitancy detection and determine which approach is best suited for our specific research objectives.

#### *3.3.4.1 Softmax*

The *softmax* function is a mathematical procedure that converts a vector of real numbers into a probability distribution whose total equals 1. Converting the output of a neural

network into probabilities that may be understood as class probabilities or probabilities of various outcomes is frequently done in machine learning and deep learning models. The formula for *softmax* function is as follows:

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{3.1}$$

Where:

$\sigma$ = Softmax.
$\vec{z}$ = input vector.
$e^{z_i}$ = standard exponential function for input vector.
$K$ = number of classes.
$e^{z_j}$ = standard exponential function for output vector.

When training machine learning models for classification tasks, the *softmax* function is commonly used in combination with a *cross-entropy* loss function. The *cross-entropy* loss function measures the difference between the actual distribution of the data and the predicted probability distribution. By minimising the cross-entropy loss, the model is able to fine-tune its parameters to make accurate predictions. This helps the model learn how to produce precise outputs during training.

### 3.3.5 Performance evaluation

In the final step of the methodology, the performance of the model must be scrutinised. As our main objective of this research is to evaluate how our proposed model performs in mono-platform versus multi-platform settings. Hence, the performance of the model needs to be evaluated and compared separately using platform specific model and multi-platform model.

The evaluation of machine learning models can be performed through various methods. In this study, the model employed is a classification model that aims to classify social media comments as either vaccine hesitant or not. A variety of evaluation criteria were employed to assess the classification model, including accuracy, precision, recall, and F1-score. These metrics are widely used in classification tasks to measure the effectiveness of the model in predicting the correct class labels. The accuracy metric measures the overall correctness of the model's predictions. Precision and recall metrics evaluate the precision and completeness of the model's predictions, respectively. Finally, the F1-score metric is a harmonic mean of precision and recall and provides a balanced measure of the model's performance.

#### *3.3.5.1 Confusion Matrix*

In a classification task, model testing is done on a testing dataset to measure the model's performance. The classification task's output is divided into various categories, which are: True positive (TP), True Negative (TN), False positive (FP) and False negative (FN) (H. Han et al., 2005). Figure 3.3 represents a confusion matrix table. A confusion matrix is a useful tool for assessing the effectiveness of a classification model. It presents a table that outlines the number of TP, TN, FP and FN predictions made by the model. TP reflects the number of instances correctly identified as positive, TN shows the number of instances correctly identified as negative, FP represents the number of negative instances incorrectly labelled as positive, and FN reflects the number of positive instances wrongly labelled as negative. These values can be used to calculate various evaluation metrics such as accuracy, precision, recall, and F1-score.

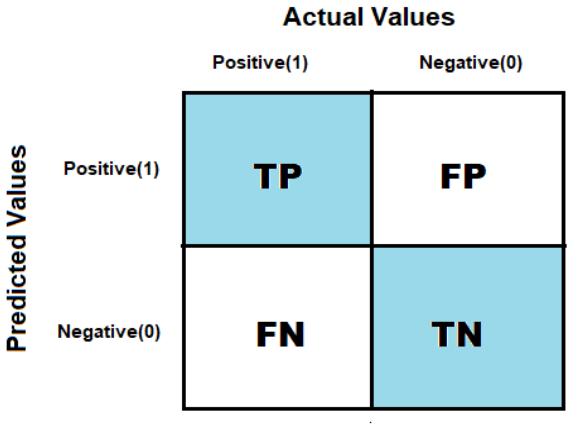



Figure 3.3 Confusion matrix

In our vaccine hesitancy detection model, the testing dataset is labelled as 1 or 0. 1 represents the comments that are vaccine hesitant while '0' represents not vaccine hesitant.

### *3.3.5.2 Accuracy Rate*

Accuracy is a commonly used metric to evaluate the performance of a model. It represents the proportion of correct predictions made by the model on the test dataset, out of all the predictions. While a high accuracy rate generally indicates good model performance, it can be misleading in the case of imbalanced class distribution in the test dataset. For instance, if we have 100 test instances with 90 instances belonging to the positive class and the model predicts all instances as positive, the accuracy would be 90%, even though the model is not actually learning anything meaningful. The formula for accuracy rate is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.2}$$

#### *3.3.5.3 Precision and Recall*

Precision is a metric used to measure the model's ability to correctly predict positive values. It is the ratio of true positive predictions to the total number of positive predictions made by the model. The precision score is calculated using the following formula:

$$Precision = \frac{TP}{TP + FP} \tag{3.3}$$

A high precision score indicates that the model makes very few false positive predictions. In other words, when the model predicts a positive value, it is highly likely to be correct. However, a high precision score does not necessarily mean that the model has a good overall performance, as it may miss many true positive predictions.

Recall is a metric used to evaluate the completeness of the classification model. It measures the proportion of actual positive samples that are correctly identified by the model as positive (true positive). In other words, recall indicates the ability of the model to detect positive instances correctly. The formula for recall is:

$$Recall = \frac{TP}{TP + FN} \tag{3.4}$$

where true positives are the number of instances accurately classified as positive, and false negatives are the instances that are actually positive but incorrectly classified as negative. High recall indicates that the model is effectively identifying most of the positive instances. However, it does not take into account the false positives, which may lead to lower precision.

#### *3.3.5.4 F1-score*

To fully assess a model's effectiveness, we must consider both precision and recall. Unfortunately, precision and recall are frequently at odds. In other words, increasing precision typically decreases recall and vice versa. The F1-score balances the difference between recall and precision and checks to determine if the outcomes line up. The greater the F1-score, the better the algorithm is functioning, and the recall and precision are in agreement.

$$F = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3.5}$$

## 3.4 SUMMARY

Machine learning is a trial-and-error approach where fine tuning the model is required based on the model's performance. The data collection and preprocessing tasks must be carefully conducted so that the right data is fed into the model. Model evaluation is crucial to understand the performance of the model in terms of its predictive accuracy, ability to generalize, and overall quality. Meanwhile, evaluation metrics gauge the model's effectiveness in comparison to other models or algorithms.

# CHAPTER FOUR
# EXPERIMENTAL SETUP

## 4.1 INTRODUCTION

In Chapter 3, the methodology of this research has been introduced. This chapter presents the experimental setup employed for our research. Section 4.2 details the process of dataset sampling, Section 4.3 focuses on data analysis while Section 4.4 outlines the modelling strategy for training the models. Lastly, Section 4.5 discusses the evaluation metrics used to assess the performance of the models.

## 4.2 DATA SAMPLING

The dataset used in this study was collected and aggregated from the sources mentioned in Section 3.3.1. Once the data has been aggregated, it underwent several pre-processing techniques, as described in Section 3.3.2. A snippet of the pre-processed dataset is shown in Table 4.1, which includes the comments or tweets from the social media in the *comments* column. The *sentiment* column indicates the public outlook towards vaccination, with a label of 1 for negative sentiment and 0 for positive sentiment. The last column named *platform* represents the source of social media from where the data was obtained.

Table 4.1 Dataset Sample

| **Comments** | **Sentiment** | **Platform** |
|---|---|---|
| i¡¯m proud lakelandcommcol good turnout covid19 vaccine today got moderna | 0 | Twitter |

| | | |
|---|---|---|
| germany suspends use oxfordastrazeneca vaccine reason looks distracting blooming useless vaccine roll good luck getting people vaccinated calling doubt safety vaccine idiots | 1 | Twitter |
| it’s normal questions vaccines answers common questions | 0 | Facebook |
| vaccines protect 100 hospitalization death clinical trials percent top immune system offers default we’re still trusting immune system don’t want tiny minority healthy people gets super bad hospital weeks that’s motivation getting vaccine you’re concerned “long term effects” trust unassisted immune system job done guess | 0 | Reddit |
| masks dont help lets wear anyway vaccines dont help lets anyway refugees destroy wealth lets invite anyway inflation wont save dying economy lets anyway civilization killing planet lets keep anyway ignorance infamous pairing stupid evil | 1 | Reddit |
| mainstream media admits engaging censorship response “antivaccine” ie proinformed consent activists former hhs head kathleen sebelius explicitly called tenure | 1 | Facebook |
| flu vaccine cause alzheimers many brain damages memory losses dr weil | 1 | Facebook |
| healthy people worldwide dying experimental vaccines may think cant happen healthy boy 13 dies sleep receiving second covid vaccine dose via westjournalism | 1 | Twitter |
| shot doesnt work keep getting anyway | 1 | Facebook |

Following the data pre-processing stage, the dataset was partitioned into two distinct subsets: training and testing, with a ratio of 90:10, respectively which is 173k and

19k. The testing subset was exclusively used for multi-platform performance evaluation, while the training subset was divided into three partitions: training set (80%) or 138.4k, validation set (10%) or 17.3k, and testing set (10%) or 17.3k for the purpose of mono-platform model training and evaluation. The dataset sampling strategy is depicted in Figure 4.1.

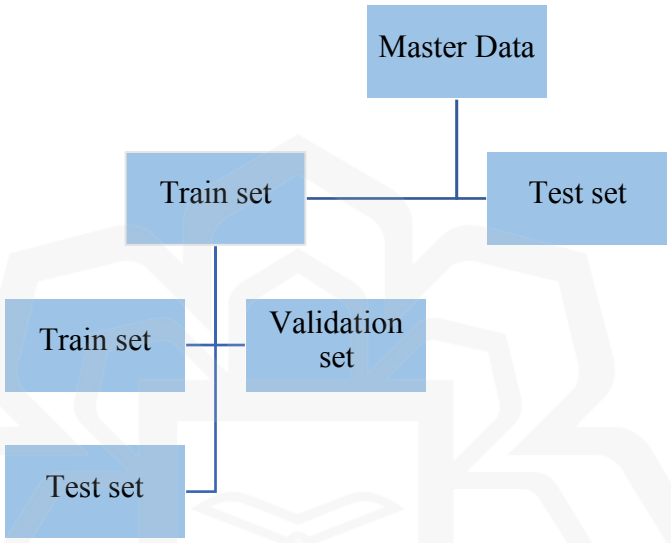



Figure 4.1 Dataset distribution for modeling

Firstly, the dataset is uploaded to google drive and later synced with Google Colaboratory. Google Colaboratory is a product of Google research. It is used for running any Python code in the browser. For seamless training of our model, subscription to Google Colab Pro+ was activated which allows the use of a high-performance GPU for executing python codes involving large datasets.

## 4.3 DATA ANALYSIS

The final accumulated dataset after preprocessing contains 193,023 annotated data. Figure 4.2 represents the distribution of data by social media platforms. Reddit, Facebook and Twitter contributes 89,470, 66,822 and 36,731 data respectively. More data was gathered

from Reddit as compared to Facebook and Twitter. However, each platform holds more than 30k dataset which is sufficient for any deep learning models.

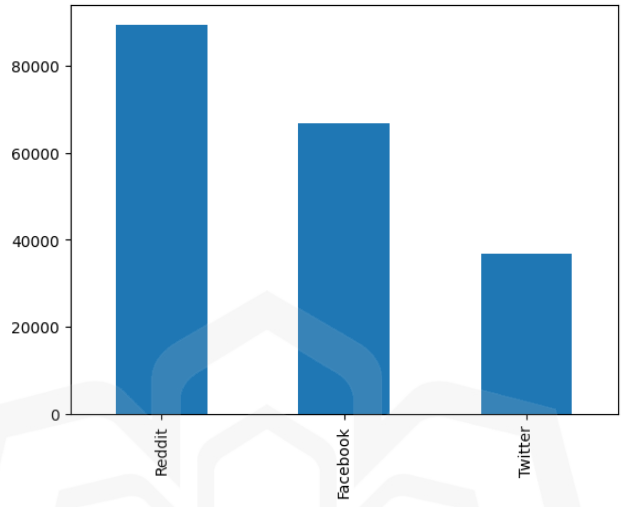



Figure 4.2 Distribution of data by social media

Figure 4.3 shows the number of positive and negative sentiments in the dataset. '0' represents positive sentiment and '1' represents negative sentiment of the data. Although there is a slightly higher number of positive sentiments, the total number of positive and negative sentiments does not significantly differ, making the dataset almost balanced.

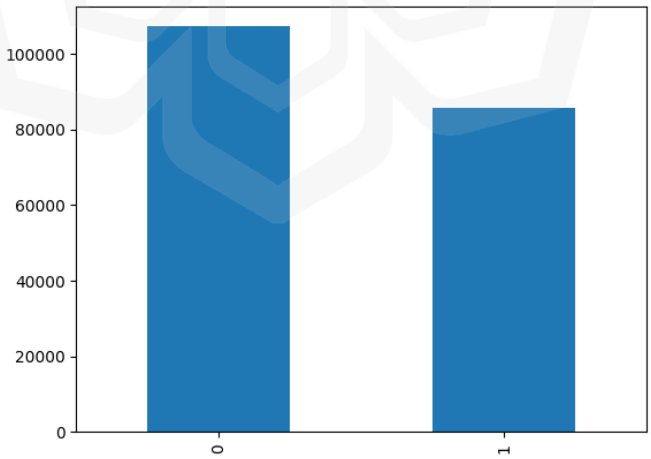



Figure 4.3 Sentiment distribution

Figure 4.4 depicts the word cloud visualisation of our Facebook, Reddit and Twitter data. It shows that there are certain words that are frequently used by the users when talking about vaccine hesitancy.

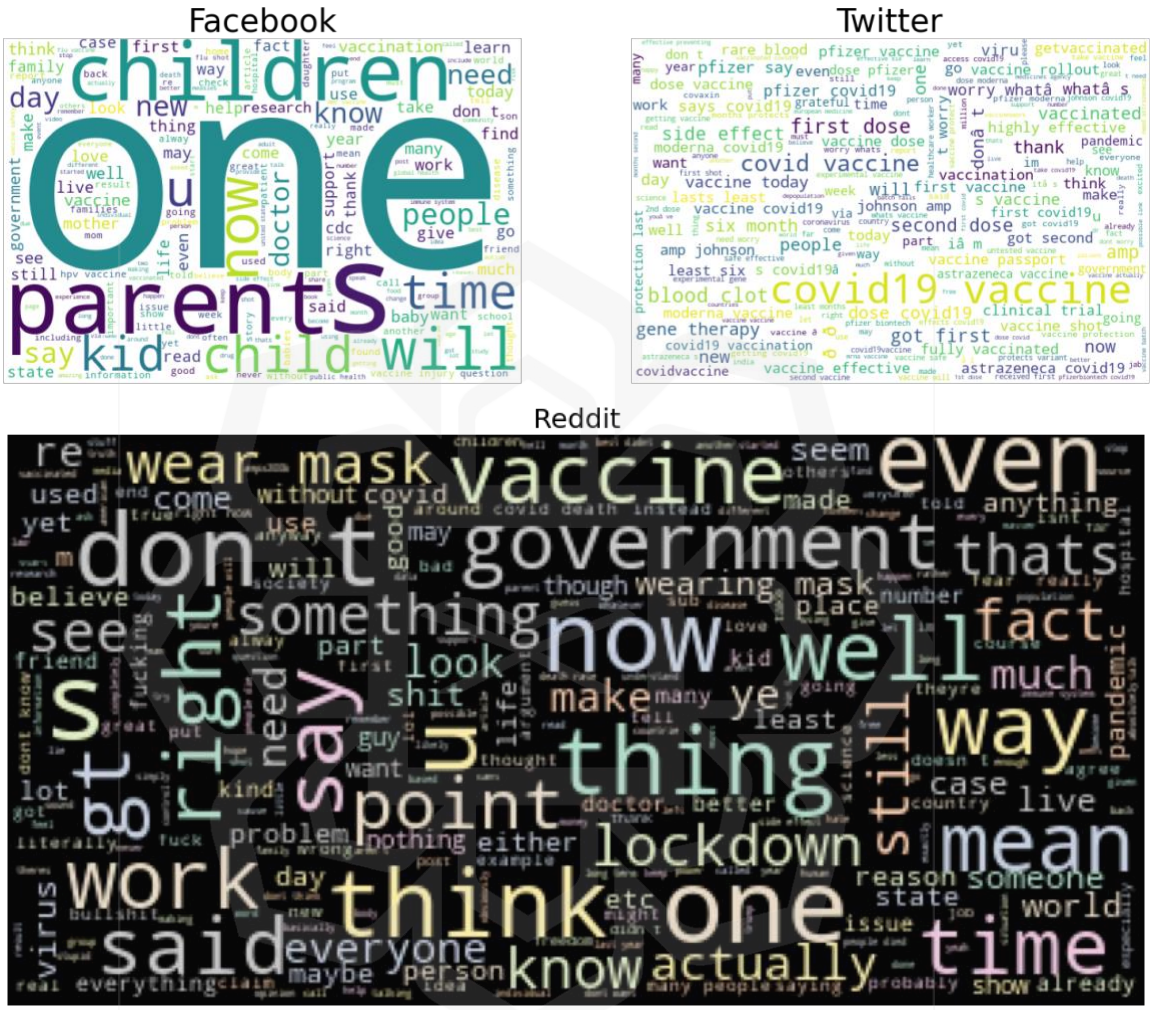



Figure 4.4 Wordcloud visualisation on vaccine hesitancy of social media dataset

Figure 4.5, illustrates the most frequently used positive and negative words in the dataset related to vaccine hesitancy. The analysis reveals that social media users tend to use the words “covid19”, “vaccine”, “thank”, “even”, “right” and “people” when expressing a positive outlook regarding vaccination. Conversely, the words “one”, “don’t”, “time”, “think”, and “now” are most widely used when expressing a negative outlook towards vaccination on social media. This information can be used to understand the most common

perspectives and opinions regarding vaccine hesitancy in social media, which can help inform public health campaigns and interventions.


Figure 4.5 Positive and negative sentiment visualisation

Figure 4.6 presents the top fifteen (15) negative words used by social media users to express their outlook towards vaccination on Facebook. The analysis shows that words such as “children”, “health”, “vaccination”, and “medical” are frequently used to express negative sentiments. Other words such as “one”, “us” and “know” also appear in the list. These findings suggest that users expressed concerns about the safety and efficacy of vaccination and the impact on their health and well-being. Additionally, the use of words such as “dr” and “parents” implies that medical professionals and parents are actively involved in the discussion around vaccination hesitancy on Facebook. The findings provide insights into the factors that influence public perception and attitudes towards vaccination on social media.

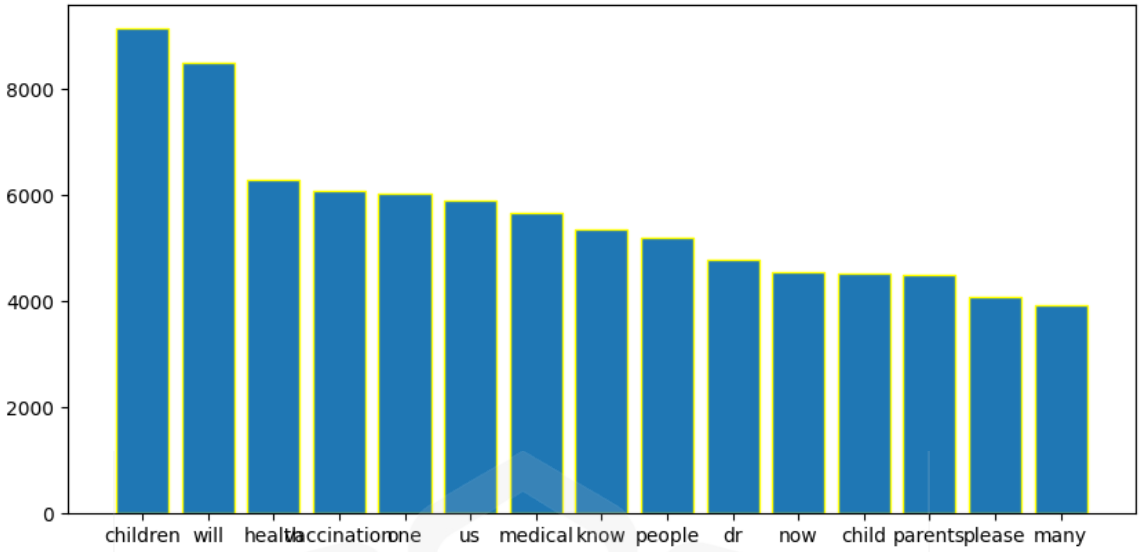



Figure 4.6 Most commonly used words in Facebook for negative outlook

Figure 4.7 presents a word cloud visualisation of the fifteen (15) most frequently used words related to negative sentiments towards vaccine hesitancy in Reddit. The most common words include “don’t”, “will”, “even”, “one”, “vaccine”, “death”, “know”, “think”, “virus”, “us”, “now”, “masks”, “im”, “mask”, and “deaths”.

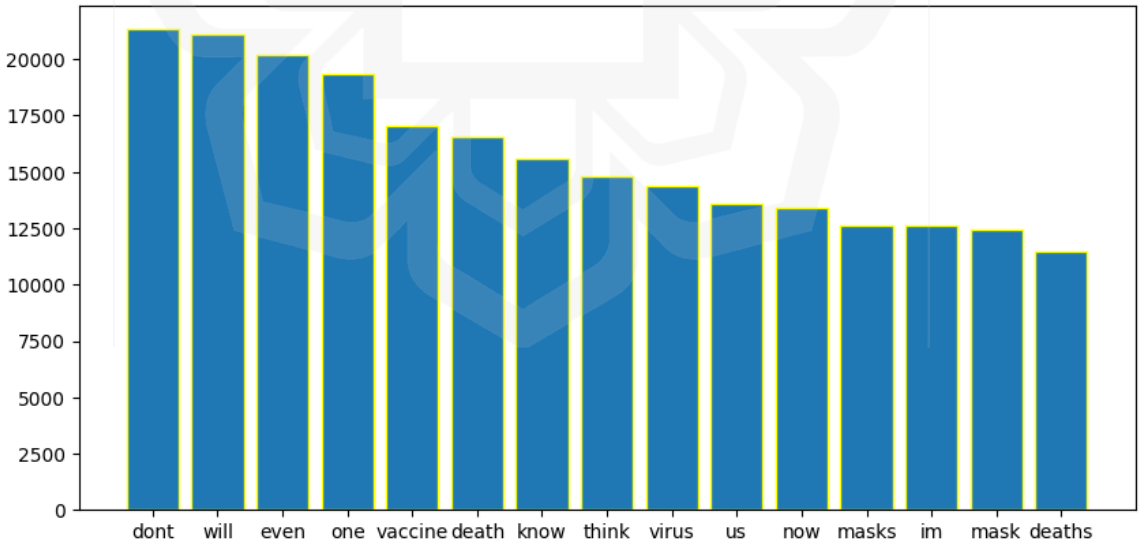



Figure 4.7 Most commonly used words in Reddit for negative outlook

In Figure 4.8, the top fifteen (15) words frequently used in expressing negative outlook towards vaccination on Twitter are shown. The words are ranked based on their frequency of usage in the dataset. These words are: “vaccines”, “blood”, “people”, “experimental”,

“covid”, “astrazeneca”, “clots”, “rare”, “take”, “will”, “amp”, “ene”, “therapy”, “says”, and “us”.

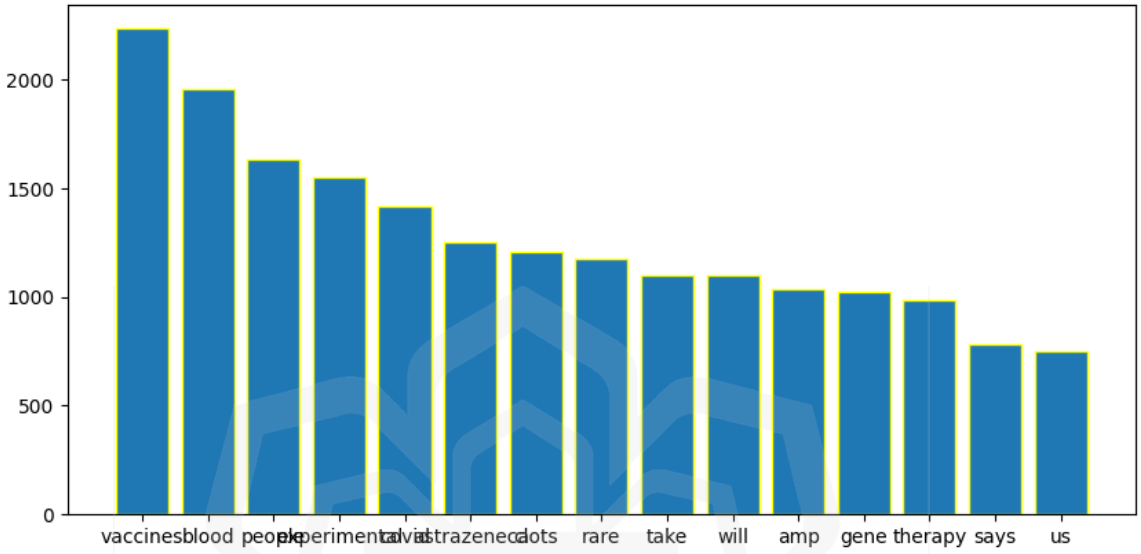



Figure 4.8 Most used words in Twitter for negative outlook

## 4.4 DATA MODELLING

As explained in Section 2.2.4.1, there are different types of BERT models available from Google such as BERT-base, BERT-large and BERT-small. BERT-large is computationally expensive to train. Hence, in this research BERT-base and BERT-small model are chosen which contains the parameters shown in Table 4.2:

Table 4.2 Parameters of BERT-base and BERT-small

| Model | Transformer Block (L) | Hidden Size (H) | Attention Heads (A) |
|---|---|---|---|
| BERT-base | 12 | 768 | 12 |
| BERT-small | 8 | 512 | 8 |

The pretrained model and the preprocessor were downloaded from TensorFlow hub. The preprocessor processes the input data accordingly to be fed into the model. The model summary is presented in Figure 4.9.

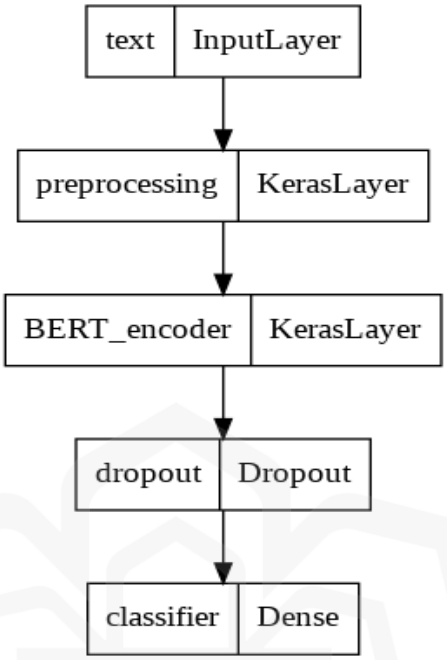



Figure 4.9 Model architecture

In the model architecture, the input layer is responsible for taking in the raw training and validation dataset. The data is then pre-processed in the Keraslayer, which is built on top of the TensorFlow handle pre-processing mechanism. This layer transforms the data for the subsequent BERT layer. After that, a dropout layer is added to prevent overfitting, followed by a dense layer or classifier layer.

Table 4.3 represents the hyperparameters used during the model training. The loss function used in this model is Binary_Crossentropy, which calculates the cross-entropy loss between the true and predicted labels. Initially, the total number of epochs was set to 20 and the learning rate 2e-5. AdamW was used as the model optimiser. Batch size was set to 32.

Table 4.3 BERT model hyperparameter tuning

| **Hyperparameter** | **Name** |
|---|---|
| Loss Function | BinaryCrossentropy |
| Epochs | 20 |
| Learning rate | 2e-5 |
| Optimiser | AdamW |
| Batch size | 32 |
| Regularisation technique | Early Stopping |

The training strategy of the model includes the adoption of an early stopping mechanism, which is a regularisation technique aimed at preventing the model from overfitting. The implementation of this model was done using TensorFlow deep learning framework. The implementation of the SVM and LR models is a straightforward process that utilises the Scikit-learn machine learning library for training and testing.

To summarise, a total of sixteen models were trained and built for the study. For each social media platform (Facebook, Twitter, and Reddit), four models were trained, including BERT-base, BERT-small, SVM, and LR. In addition, a multi-platform model was developed using data from all three platforms. The mono-platform models were trained using data only from their respective platforms, while the multi-platform model was trained using data from all three platforms. Each model was evaluated twice, once using mono-platform data and again using multi-platform data. Figure 4.10 represents an example of mono-platform model and multi-platform model.

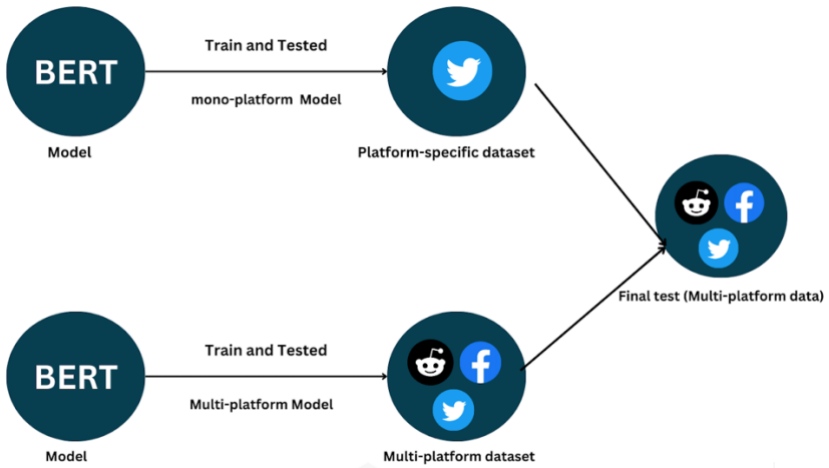



Figure 4.10 Example of mono-platform and multi-platform model

## 4.5 EVALUATION METRICS

The performance of the models will be evaluated using four popular metrics called accuracy, recall, precision and F1-score. Accuracy, recall, precision, and F1-score are four popular metrics for evaluating the performance of machine learning models, especially classification models. These metrics can be used to assess how well a model is able to identify and classify different classes of data. Table 4.4 presents the summary of model evaluation metrics.

Table 4.4 Model evaluation metrics

| | |
|---|---|
| Accuracy | Percentage of predictions that are correct |
| Recall | Percentage of actual positives that are correctly predicted |
| Precision | Percentage of predicted positives that are actually positive |
| F1-score | Harmonic mean of precision and recall |

# CHAPTER FIVE
# RESULTS

## 5.1 INTRODUCTION

In Chapter 4, an exposition on the experimental setup, encompassing data sampling, modeling, and evaluation metrics employed in this research, has been meticulously elucidated. This chapter presents an analysis of the performance of the proposed model in both mono-platform and multi-platform settings, based on the evaluation metrics discussed in the previous section. Specifically, Section 5.2 focuses on the performance of the BERT-small and BERT-base models during training, while Section 5.3 evaluates the performance of all models using accuracy, precision, recall, and F1-score. Finally, Section 5.4 and Section 5.5 provides the discussion and summary of the performance of all the models.

## 5.2 TRAINING PERFORMANCE

In this section, we discuss the performance of BERT-base and BERT-small models during the training process. Figure 5.1 and Figure 5.2 show the training and validation accuracy of the BERT-base and BERT-small models respectively. Training accuracy and validation accuracy are two popular measures used in machine learning to assess a model's performance. The degree to which a model fits the data it was trained on is known as training accuracy. It is determined by comparing the training set's actual labels to the model's predicted labels. The degree to which a model generalises to fresh, unexplored data is called validation accuracy. It is determined by making a comparison between the model's predicted labels and the actual labels in a separate validation set that should not overlap the training set. Figure 5.1 indicates an increase in accuracy over the epochs, demonstrating that the model was learning during training.

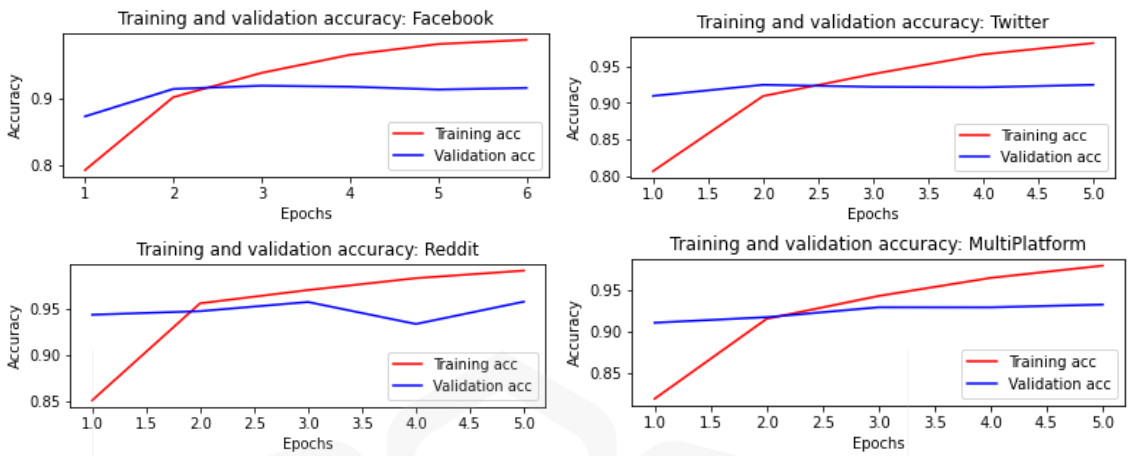



Figure 5.1 Training vs validation accuracy of BERT-base

Similarly, Figure 5.2 shows that the BERT-small model was trained well, with increasing training and validation accuracy over the epochs.

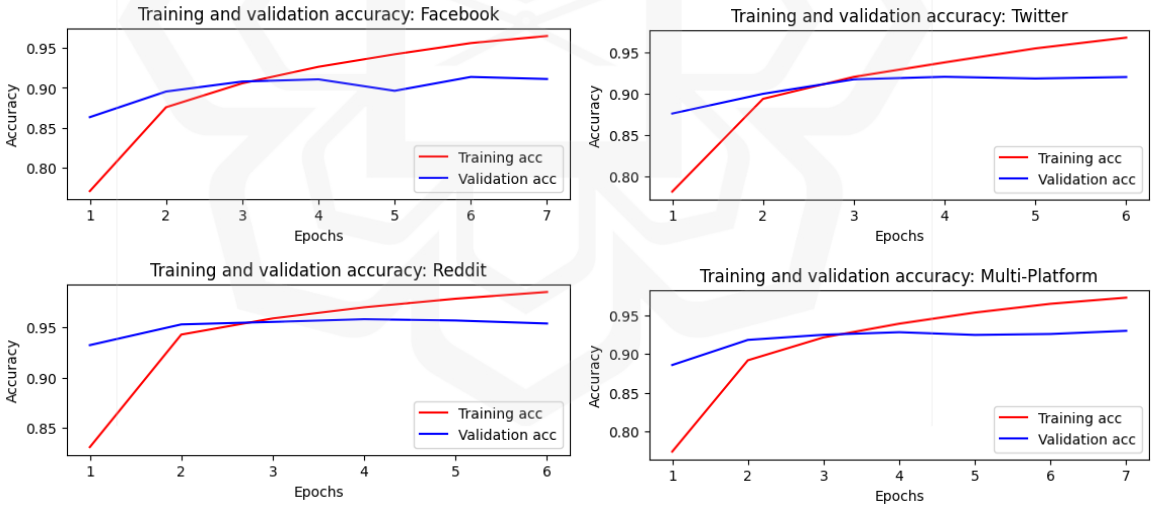



Figure 5.2 Training vs validation accuracy of BERT-small

In addition to training and validation accuracy, training and validation loss are also commonly used metrics to evaluate the performance of a machine learning model. Training loss gauges a model's proficiency at minimising the discrepancy between the output anticipated and the output obtained from the training set. For each training example, the difference between the anticipated and actual output is measured, and the average of all the

differences is then used to calculate the result. On a distinct validation set, which should not overlap with the training set, validation loss measures the discrepancy between the projected output and the actual output. Validation loss is computed the same way as training loss.

Figure 5.3 and Figure 5.4 depict the training and validation loss of the BERT-base and BERT-small models respectively. The training and validation loss of the BERT-base and BERT-small models both decreased over the course of training. This shows that over time, both models were successfully adapting and developing.

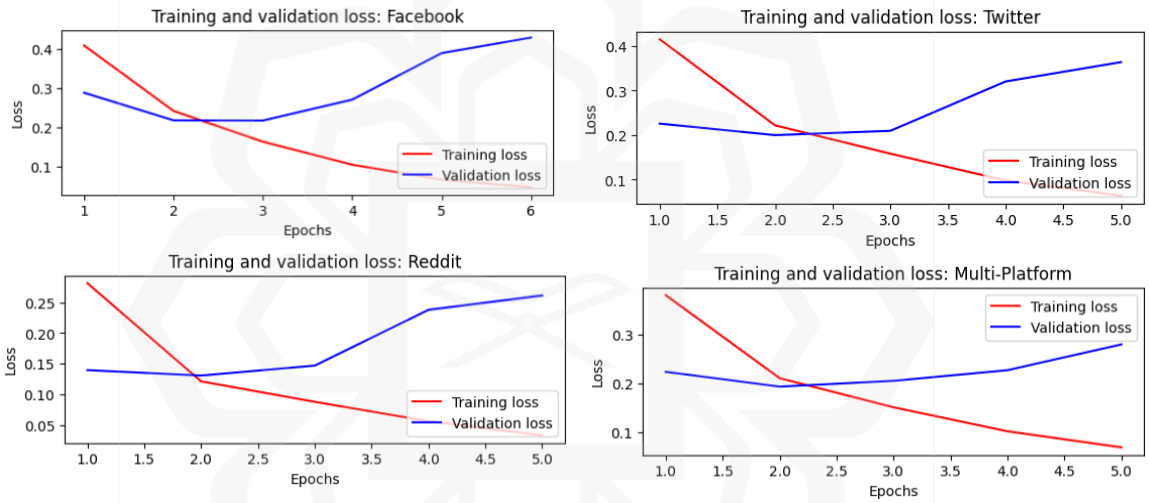



Figure 5.3 Training vs validation loss of BERT-base

It is noticeable that after a certain number of epochs, the validation loss started to increase. The regularisation technique called *early stopping* mechanism solves this problem and stops the models from overfitting. This allowed us to effectively reduce the effects of overfitting and guarantee the generalizability of the models by terminating the training process when the validation loss started to increase.


Figure 5.4 Training vs validation loss of BERT-small

The figures presented herein indicate that the hyperparameters utilised in the deep learning-based models were appropriately chosen, leading to successful learning without exhibiting overfitting. Specifically, both the BERT-small and BERT-base models attained their optimal levels of accuracy after a relatively brief period of training, typically between 5 to 7 epochs.

## 5.3 PERFORMANCE EVALUATION

This section is focused on evaluating how well the models perform using various evaluation metrics, such as accuracy, precision, recall, and F1 score. The succeeding subsections provide a comprehensive overview of each model's performances, including a summary of each performance metric using tables. The study employed four different models: BERT-small, BERT-base, SVM, and LR, and two different approaches to evaluate their performances. The first approach involved training and testing each model on mono-platform data, which means data collected from a specific social media platform, such as Twitter, Facebook, or Reddit. Additionally, each model was evaluated on multi-platform data, which is a combination of data from all three social media platforms. The second approach involved training and testing all models solely on multi-platform data to assess their ability to generalize across different social media platforms.

### 5.3.1 Performance of mono-platform model

#### *5.3.1.1 Accuracy*

Table 5.1 shows the performance of different models in terms of accuracy trained with mono-platform data. The models include BERT-small, BERT-base, SVM, and LR. Each model's performance was tested on both mono-platform data and multi-platform data. The results indicate that all models performed well when tested with the same platform they were trained on. However, the accuracy scores significantly decreased when the models were tested on multi-platform test data.

Table 5.1 Accuracy scores of different machine learning models trained with mono-platform data

| Dataset | Model | Accuracy Score (Tested on mono-platform data) | Accuracy Score (Tested on multi-platform data) |
|---|---|---|---|
| Facebook | BERT-small | 0.91 | 0.67 |
| | BERT-base | 0.91 | 0.68 |
| | SVM | 0.90 | 0.67 |
| | LR | 0.89 | 0.67 |
| Twitter | BERT-small | 0.92 | 0.64 |
| | BERT-base | 0.92 | 0.68 |
| | SVM | 0.91 | 0.67 |
| | LR | 0.90 | 0.68 |
| Reddit | BERT-small | 0.96 | 0.81 |
| | BERT-base | 0.96 | 0.81 |
| | SVM | 0.97 | 0.80 |
| | LR | 0.96 | 0.80 |

#### *5.3.1.2 Recall*

Table 5.2 shows the recall rates of different models when tested with mono-platform data and multi-platform data. The results demonstrate that all models achieved a recall rate of at least 0.87 when tested on mono-platform data. However, the recall rate dropped by at least 20% for all models when tested with multi-platform data. Specifically, for the BERT-base model which was trained with Facebook data, the recall rate dropped from 0.91 when tested with multi-platform data. Similar patterns were observed for other models as well.

Table 5.2 Recall values of different machine learning models trained with mono-platform data

| Dataset | Model | Recall Value (Tested on mono-platform data) | Recall Value (Tested on multi-platform data) |
|---|---|---|---|
| Facebook | BERT-small | 0.91 | 0.69 |
| | BERT-base | 0.91 | 0.70 |
| | SVM | 0.90 | 0.70 |
| | LR | 0.89 | 0.69 |
| Twitter | BERT- small | 0.91 | 0.60 |
| | BERT-base | 0.90 | 0.66 |
| | SVM | 0.87 | 0.65 |
| | LR | 0.86 | 0.66 |
| Reddit | BERT-small | 0.96 | 0.80 |
| | BERT-base | 0.96 | 0.79 |
| | SVM | 0.97 | 0.79 |
| | LR | 0.97 | 0.79 |

#### *5.3.1.3 Precision*

Table 5.3 shows that in terms of precision rate, the platform specific models performed well, achieving a precision rate of over 0.90 when tested on mono-platform data. However, when tested with multi-platform data, the precision rate dropped. Specifically, the precision rate for the SVM model trained with Reddit data decreased from 0.97 to 0.82.

Table 5.3 Precision values of different machine learning models trained with mono-platform data

| Dataset | Model | Precision Value (Tested on mono-platform data) | Precision Value (Tested on multi-platform data) |
|---|---|---|---|
| Facebook | BERT- small | 0.91 | 0.71 |
| | BERT-base | 0.91 | 0.72 |
| | SVM | 0.90 | 0.71 |
| | LR | 0.90 | 0.70 |
| Twitter | BERT- small | 0.90 | 0.66 |
| | BERT-base | 0.90 | 0.70 |
| | SVM | 0.91 | 0.69 |
| | LR | 0.91 | 0.71 |
| Reddit | BERT- small | 0.96 | 0.82 |
| | BERT-base | 0.96 | 0.82 |
| | SVM | 0.97 | 0.82 |
| | LR | 0.97 | 0.82 |

#### *5.3.1.4 F1-score*

In assessing the performance of classification models, the F1-score is a critical metric. Similar to the other performance metrics, the platform specific models demonstrated poor performance when evaluated with multi-platform data. Among all the models, the LR

model attained the lowest F1-score of 0.88 when trained with Twitter data. However, both the LR and SVM models achieved the maximum F1-score of 0.97 when trained and tested with Reddit social media data.

Table 5.4 F1-scores of different machine learning models trained with mono-platform data

| Dataset | Model | F1-score (Tested on mono-platform data) | F1-score (Tested on multi-platform data) |
|---|---|---|---|
| Facebook | BERT-small | 0.91 | 0.67 |
| | BERT-base | 0.91 | 0.65 |
| | SVM | 0.90 | 0.68 |
| | LR | 0.89 | 0.67 |
| Twitter | BERT-small | 0.90 | 0.58 |
| | BERT-base | 0.90 | 0.65 |
| | SVM | 0.89 | 0.64 |
| | LR | 0.88 | 0.65 |
| Reddit | BERT-small | 0.96 | 0.80 |
| | BERT-base | 0.96 | 0.80 |
| | SVM | 0.97 | 0.80 |
| | LR | 0.97 | 0.80 |

### 5.3.2 Performance of multi-platform model

Table 5.5 provides a summary of the performance evaluation results of the proposed multi-platform model based on accuracy, precision, recall, and F1-score. The analysis revealed that the BERT-small and BERT-base models outperformed SVM and LR models. The BERT-base model exhibited the highest F1-score of 0.93, indicating that it can accurately predict vaccine hesitancy across multiple social media platforms. As illustrated in Table

5.5, the overall performance score of the multi-platform model exceeded 0.90, which is substantially better than the mono-platform models, as discussed in earlier sections.

Table 5.5 Overall performance of multi-platform model

| Model | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| BERT-small | 0.92 | 0.92 | 0.92 | 0.92 |
| BERT-base | 0.93 | 0.93 | 0.93 | 0.93 |
| SVM | 0.90 | 0.90 | 0.90 | 0.90 |
| LR | 0.90 | 0.90 | 0.90 | 0.90 |

## 5.4 DISCUSSION

Figure 5.5 provides a graphical representation of the decrease in performance of mono-platform models when tested with multi-platform settings. These findings highlight the importance of training models with multi-platform data for improved performance in vaccine hesitancy detection across various social media platforms.

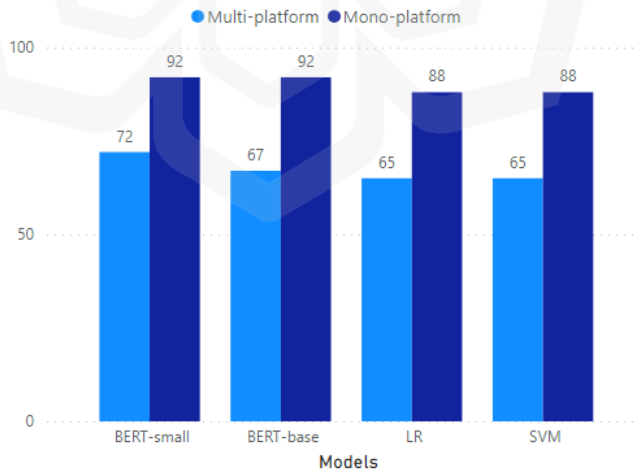



Figure 5.5 Performance of models trained with mono-platform data (Tested on mono-platform and multi-platform data)

In addition, Figure 5.6 illustrates the superior performance of the model trained with multi-platform data as compared to the mono-platform models trained with Reddit, Facebook, and Twitter data.

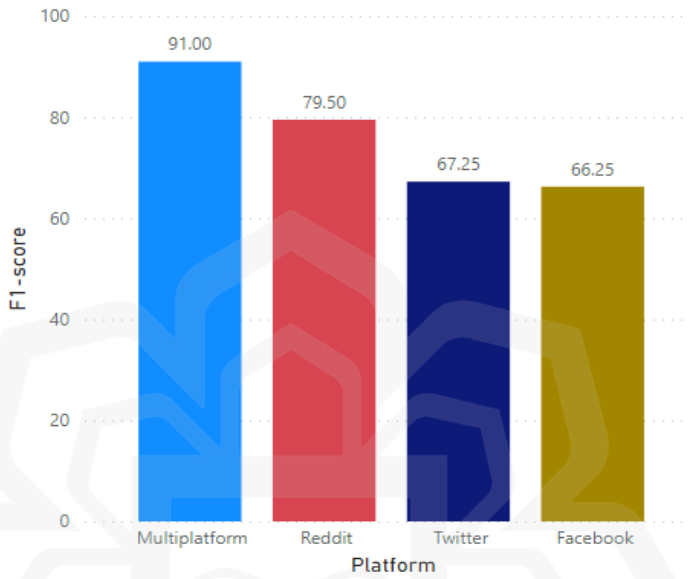



Figure 5.6 Model's performance by platforms

As stated in Section 2.3.4, Lemmens et al., (2021) has conducted a research that involved using data from Twitter and Facebook to develop a BERT model for detecting vaccine hesitancy in Dutch. The model achieved an F1-score of 0.73. In contrast, this research study used the BERT model and trained it with data from Facebook, Twitter, and Reddit. The results of this research were significantly better, with an F1-score of 0.93. The results from Figure 5.7 show that the BERT-base model achieved the best performance in multi-platform vaccine hesitancy detection, followed by the BERT-small model while both the SVM and LR models exhibited similar performances with slightly lower scores than the BERT models.

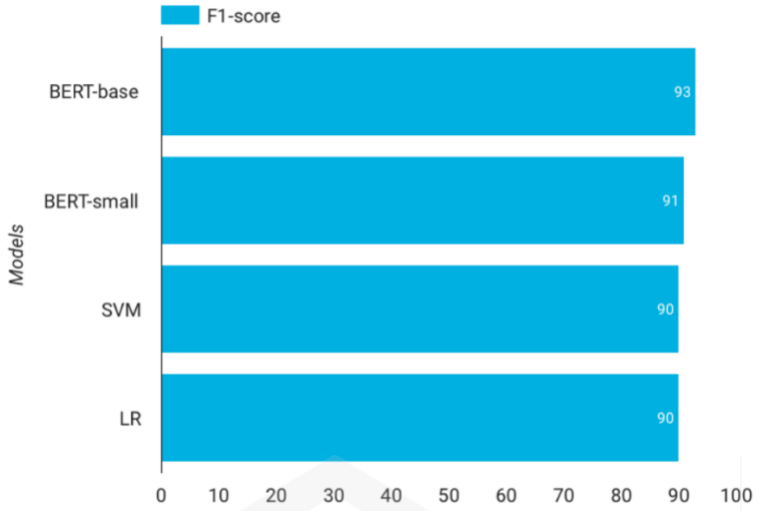



Figure 5.7 Performance of the best model based on F1-score

Based on the findings of the study, it can be inferred that the use of mono-platform models for detecting vaccine hesitancy across multiple social media platforms is not advisable due to their poor performances. Conversely, the research demonstrated that the BERT model exhibited superior performance when compared against other machine learning models in detecting vaccine hesitancy when trained using data from various social media platforms. The findings of this research also demonstrated superior performance compared to the existing study by Lemmens et. al (2021) in detecting vaccine hesitancy using multi-platform data, as evidenced by the detection accuracy results. Hence, it can be concluded that the BERT model is a more effective approach for detecting vaccine hesitancy across multiple social media platforms.

## 5.5 SUMMARY

The analysis of the performance metrics revealed that the mono-platform models exhibited poor performances when tested with multi-platform data. The findings indicated that a model trained with multi-platform data outperformed a model trained with mono-platform data at least by 15%. Therefore, the model developed with Facebook data for detecting vaccine hesitancy exhibited poor performance when applied to detect vaccine hesitancy on Twitter or Reddit. Similarly, the performance of any mono-platform model, such as Twitter

or Reddit, also suffered. The only model that performed well in detecting vaccine hesitancy is the model trained with multi-platform data.

# CHAPTER SIX
# CONCLUSION

## 6.1 INTRODUCTION

In Chapter 5, an analysis has been conducted concerning the performance of models in both mono and multi-platform settings. This chapter provides a comprehensive discussion on the conclusion, limitations, and future work related to the research on vaccine hesitancy detection using multiple social media platforms. Section 6.2 presents the research conclusion and evaluates the extent to which the research objectives have been achieved. In Section 6.3, the chapter addresses the limitations of the research and proposes future directions for further investigation. Specifically, the section identifies the potential limitations and challenges faced in the current study and provides recommendations for addressing these shortcomings. The chapter suggests that future research could expand the dataset to include data from other social media platforms such as YouTube and Instagram, explore alternative machine learning and deep learning models, and introduce a multi-classification model to improve the accuracy and detection of vaccine hesitancy.

## 6.2 CONCLUSION

Vaccine hesitancy is a pressing issue in various social media platforms. To effectively address this issue, it is crucial to develop a robust vaccine hesitancy detection framework that incorporates multi-platform data. In this research, a framework for vaccine hesitancy detection that employed the BERT model and integrated data from multiple social media platforms had been proposed. Our proposed model demonstrated a significant performance improvement compared to models that utilised mono-platform data.

The first objective of this research is *to establish a consolidated dataset from multiple social media sources for use in vaccine hesitancy detection*. Within the timeline

of this research study, no prior research has been reported on the use of consolidated and published data from multiple social media platforms for this purpose. Accordingly, data were consolidated from three prominent social media platforms, namely Twitter, Facebook, and Reddit. In line with our first objective, we plan to make this dataset publicly available through the Github platform as a contribution to the research community for vaccine hesitancy detection.

The second objective of the research is *to evaluate the effectiveness of using mono-platform versus multi-platform vaccine hesitancy data on the performance of different machine learning models.* To achieve this objective, four models were trained and tested: BERT-base, BERT-small, SVM, and LR. In the previous chapter, the performances of these models were compared, and it was found that all models trained with single platforms exhibited poor performance when tested with data from multiple social media platforms. In contrast, models trained with multi-platform data showed a performance increase of at least 15%. This will serve as valuable guidance for future researchers, encouraging them to contemplate the integration of data from multiple platforms when conducting vaccine hesitancy detection studies.

The third objective of the research is *to apply a transfer learning method using BERT in vaccine hesitancy detection.* We trained and tested the BERT model using multi-platform data. The F1-score of the BERT model for vaccine hesitancy detection from multiple social media platforms is 0.93. Indeed, the BERT-based transfer learning method performed the best as compared to other machine learning methods such as SVM and LR. Furthermore, a model with F1-score of 0.93 is considered a good model for implementation in production settings. Hence, it can be concluded that the BERT-based transfer learning method for vaccine hesitancy detection is one of the best approaches available. This will facilitate future researchers in conducting a comprehensive exploration of the efficiency and capabilities of the BERT model.

## 6.3 LIMITATION AND FUTURE WORK

This research work on vaccine hesitancy detection for multiple social media platforms using BERT has several limitations. These limitations are highlighted below:

- *Limitation of the dataset*: the dataset used in this research was limited to data collected from Facebook, Twitter, and Reddit. Other social media platforms, such as Instagram and Youtube, were not included, which may limit the generalizability of the findings.

- *Limitation of the models*: The study only utilised four machine learning models, including BERT-base, BERT-small, SVM, and LR. Other models, such as BERT-large, LSTM, decision tree, and random forest, were not included due to computational constraints and the nature of the research.

- *Limitation of the number of classes*: The study was limited to binary classification of vaccine hesitancy detection, which may not provide a comprehensive understanding of the nuances of vaccine hesitancy sentiment.

Based on the limitations discussed earlier, it is suggested that the following future works will be carried out by other researchers for vaccine hesitancy detection using multiple social media platforms.

- *Incorporation of data from other social media platforms*: YouTube and Instagram dataset can be incorporated into the consolidated dataset to extend the number of platforms as well as to improve the model's performance.

- *Usage of other models*: Other machine learning and deep learning models can be trained and tested to evaluate their effectiveness in vaccine hesitancy detection.

- *Usage of multi-classification*: Future studies can introduce a multi-classification model that classifies vaccine hesitant text or comments as positive, negative,

neutral, or abusive for better accuracy and detection technique. These future directions will enhance the understanding on vaccine hesitancy and inform effective communication strategies to address this pressing public health issue.

# REFERENCES


Alammar, J. (n.d.). The illustrated transformer. The Illustrated Transformer – Jay Alammar – Visualizing machine learning one concept at a time. Retrieved July 4, 2022, from *http://jalammar.github.io/illustrated-transformer/*

Alamoodi, A. H., Zaidan, B. B., Al-Masawa, M., Taresh, S. M., Noman, S., Ahmaro, I. Y., ... & Salahaldin, A. (2021). Multi-perspectives systematic review on the applications of sentiment analysis for vaccine hesitancy. *Computers in Biology and Medicine, 139*, 104957.

Argyris, Y. A., Monu, K., Tan, P.-N., Aarts, C., Jiang, F., & Wiseley, K. A. (2021). Using Machine Learning to Compare Provaccine and Antivaccine Discourse Among the Public on Social Media: Algorithm Development Study. *JMIR Public Health and Surveillance*, *7*(6), e23105. https://doi.org/10.2196/23105

Bari, A., Heymann, M., Cohen, R. J., Zhao, R., Szabo, L., Apas Vasandani, S., ... & Coffee, M. (2022). Exploring Coronavirus Disease 2019 Vaccine Hesitancy on Twitter Using Sentiment Analysis and Natural Language Processing Algorithms. *Clinical Infectious Diseases, 74*(Supplement_3), e4-e9.

Baru, C., Institute of Electrical and Electronics Engineers, & IEEE Computer Society. (2019). Analyzing Public Outlook towards Vaccination using Twitter. *2019 IEEE International Conference on Big Data : Proceedings : Dec 9 - Dec 12, 2019, Los Angeles, CA, USA*.

Bengio, Yoshua; LeCun, Yann; Hinton, Geoffrey (2015). "Deep Learning". *Nature. 521 (7553): 436–444*.

Betsch, C., Schmid, P., Heinemeier, D., Korn, L., Holtmann, C., & Böhm, R. (2018). Beyond confidence: Development of a measure assessing the 5C psychological antecedents of vaccination. *PloS one*, *13*(12), e020860

Bloom, D. E. (2011). The value of vaccination. *In Hot topics in infection and immunity in children VII (pp. 1-8)*. Springer, New York, NY.

Burki, T. (2020). The online anti-vaccine movement in the age of COVID-19. *The Lancet Digital Health*, *2*(10), e504-e505.

Cerda, A. A., & García, L. Y. (2021). Hesitation and refusal factors in individuals' decision-making processes regarding a coronavirus disease 2019 vaccination. *Frontiers in public health*, *9*, 626852.

Cha, S.-M., Lee, S.-S., & Ko, B. (2021). Attention-Based Transfer Learning for Efficient Pneumonia Detection in Chest X-ray Images. *Applied Sciences, 11*(3), 1242. MDPI AG. Retrieved from http://dx.doi.org/10.3390/app11031242

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*, 273-297.

Cotfas, L. A., Delcea, C., & Gherai, R. (2021). COVID-19 vaccine hesitancy in the month following the start of the vaccination process. *International Journal of Environmental Research and Public Health*, 18(19), 10438.

Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research*. Pearson Education, Inc.

Delobelle, P., Winters, T., & Berendt, B. (2020). Robbert: a dutch roberta-based language model. arXiv preprint arXiv:2001.06286.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Du, J., Xu, J., Song, H., Liu, X., & Tao, C. (2017). Optimization on machine learning based approaches for sentiment analysis on HPV vaccines related tweets. *Journal of Biomedical Semantics*, *8*(1). https://doi.org/10.1186/s13326-017-0120-6

Fransiska, S., Rianto, R., & Gufroni, A. I. (2020). Sentiment Analysis Provider by. U on Google Play Store Reviews with TF-IDF and Support Vector Machine (SVM) Method. *Scientific Journal of Informatics*, *7*(2), 203-212.

Garay, J., Yap, R., & Sabellano, M. J. (2019, February). An analysis on the insights of the anti-vaccine movement from social media posts using k-means clustering algorithm and VADER sentiment analyzer. In *IOP Conference Series: Materials Science and Engineering* (Vol. 482, No. 1, p. 012043). IOP Publishing.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

Hayawi, Kadhim, Sakib Shahriar, Mohamed Adel Serhani, Ikbal Taleb, and Sujith Samuel Mathew. "ANTi-Vax: a novel Twitter dataset for COVID-19 vaccine misinformation detection." *Public Health 203* (2022): 23-30

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the *IEEE conference on computer vision and pattern recognition* (pp. 770-778).

Hossain, S. M., & Sulaiman, S. (2022). Detecting Public Outlook Towards Vaccination Using Machine Learning Approaches: A Systematic Review. In *International Conference of Reliable Information and Communication Technology* (pp. 141-150). Springer, Cham.

Jafar, A., Dambul, R., Dollah, R., Sakke, N., Mapa, M. T., & Joko, E. P. (2022). COVID-19 vaccine hesitancy in Malaysia: Exploring factors and identifying highly vulnerable groups. *PLoS One*, *17*(7), e0270868.

Johnson, N. F., Velásquez, N., Restrepo, N. J., Leahy, R., Gabriel, N., El Oud, S., ... & Lupu, Y. (2020). The online competition between pro-and anti-vaccination views. *Nature*, *582*(7811), 230-233.

Joshi, A., Dai, X., Karimi, S., Sparks, R., Paris, C., & MacIntyre, C. R. (2019). *Shot Or Not: Comparison of NLP Approaches for Vaccination Behaviour Detection*. 43–47. https://doi.org/10.18653/v1/w18-5911

Kang, Z., Catal, C., & Tekinerdogan, B. (2020). Machine learning applications in production lines: A systematic literature review. *Computers & Industrial Engineering, 149*, 106773.

Kelleher, J. D. (2019). *Deep learning*. MIT press.

Lemmens, J., Dejaeghere, T., Kreutz, T., Van Nooten, J., Markov, I., & Daelemans, W. (2021). Vaccinpraat: monitoring vaccine skepticism in Dutch Twitter and Facebook comments. *Computational Linguistics in the Netherlands Journal,* 11, 173-188.

Lemmens, J., Van Nooten, J., Kreutz, T., & Daelemans, W. (2022, October). CoNTACT: A Dutch COVID-19 Adapted BERT for Vaccine Hesitancy and Argumentation Detection. In *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 6837-6845).

Liddy, E. D. (2001). Natural language processing.

Liu, S., Li, J., & Liu, J. (2021). Leveraging transfer learning to analyze opinions, attitudes, and behavioral intentions toward COVID-19 vaccines: Social media content and temporal analysis. In *Journal of Medical Internet Research* (Vol. 23, Issue 8). JMIR Publications Inc. https://doi.org/10.2196/30251

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

MacDoanl, N. E. (2015). Vaccine hesitancy: Definition, scope and determinants. *Vaccine, 33*(34), 4161-4164.

Marcec, R., & Likic, R. (2022). Using twitter for sentiment analysis towards AstraZeneca/Oxford, Pfizer/BioNTech and Moderna COVID-19 vaccines. *Postgraduate Medical Journal, 98*(1161), 544-550

Mee, A., Homapour, E., Chiclana, F., & Engel, O. (2021). Sentiment analysis using TF–IDF weighting of UK MPs' tweets on Brexit. *Knowledge-Based Systems*, *228*, 107238.

Mejova, Y. (2009). Sentiment analysis: An overview. *University of Iowa, Computer Science Department*.

Meppelink, C. S., Hendriks, H., Trilling, D., van Weert, J. C. M., Shao, A., & Smit, E. S. (2021). Reliable or not? An automated classification of webpages about early childhood vaccination using supervised machine learning. *Patient Education and Counseling*, *104*(6), 1460–1466. https://doi.org/10.1016/j.pec.2020.11.013

Mitchell, T. M., & Mitchell, T. M. (1997). *Machine learning* (Vol. 1, No. 9). New York: McGraw-hill.

Muric, G., Wu, Y., & Ferrara, E. (2021). COVID-19 Vaccine Hesitancy on Social Media: Building a Public Twitter Data Set of Antivaccine Content, Vaccine Misinformation, and Conspiracies. *JMIR public health and surveillance*, 7(11), e30642.

Na, T., Cheng, W., Li, D., Lu, W., & Li, H. (2021). Insight from NLP analysis: COVID-19 vaccines sentiments on social media. *arXiv preprint arXiv:2106.04081*.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval, 2*(1–2), 1-135.

Piedrahita-Valdés, H., Piedrahita-Castillo, D., Bermejo-Higuera, J., Guillem-Saiz, P., Bermejo-Higuera, J. R., Guillem-Saiz, J., Sicilia-Montalvo, J. A., & Machío-Regidor, F. (2021). Vaccine hesitancy on social media: Sentiment analysis from June 2011 to April 2019. *Vaccines*, *9*(1), 1–12. https://doi.org/10.3390/vaccines9010028

Plotkin, S. (2014). History of vaccination. *Proceedings of the National Academy of Sciences,* 111(34), 12283-12287

Qorib, M., Oladunni, T., Denis, M., Ososanya, E., & Cotae, P. (2023). COVID-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset. *Expert Systems with Applications, 212*, 118715.

Ravichandiran, S. (2021). *Getting Started with Google BERT: Build and train state-of-the-art natural language processing models using BERT*. Packt Publishing Ltd.

Renee Garett, Sean D Young, Online misinformation and vaccine hesitancy, *Translational Behavioral Medicine*, Volume 11, Issue 12, December 2021, Pages 2194–2199, https://doi.org/10.1093/tbm/ibab128

Rodríguez-González, A., Tuñas, J. M., Fernandez Peces-Barba, D., Menasalvas-Ruiz, E., Jaramillo, A., Cotarelo, M., Conejo, A., Arce, A., Gil, A., Rey, U., & Carlos, J. (2020). *Creating a metamodel based on machine learning to identify the sentiment of vaccine and disease-related messages in Twitter: the MAVIS study*. https://www.ibm.com/watson/services/tone-analyzer/

Ruiz, J., Featherstone, J. D., & Barnett, G. A. (2021, January). Identifying Vaccine Hesitant Communities on Twitter and their Geolocations: A Network Approach. In *Proceedings of the 54th Hawaii international conference on system sciences* (p. 3964).

Salminen, J., Hopf, M., Chowdhury, S. A., Jung, S. G., Almerekhi, H., & Jansen, B. J. (2020). Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences, 10*(1), 1-34.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks, 61*, 85-117.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Skafle, I., Nordahl-Hansen, A., Quintana, D. S., Wynn, R., & Gabarron, E. (2022). Misinformation about COVID-19 vaccines on social media: rapid review. *Journal of medical Internet research*, *24*(8), e37367.

Sutton, R. S., & Barto, A. G. (1999). *Reinforcement learning. Journal of Cognitive Neuroscience, 11*(1), 126-134.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Taeb, M., Chi, H., & Yan, J. (2021, December). Applying Machine Learning to Analyze Anti-Vaccination on Tweets. In 2021 IEEE International Conference on Big Data (Big Data) (pp. 4426-4430). IEEE.

Thelwall, M., & Buckley, K. (2013). Topic-based sentiment analysis for the social web: The role of mood and issue-related words. *Journal of the American Society for Information Science and Technology, 64(*8), 1608-1617.

To, Q. G., To, K. G., Huynh, V. A. N., Nguyen, N. T. Q., Ngo, D. T. N., Alley, S. J., Tran, A. N. Q., Tran, A. N. P., Pham, N. T. T., Bui, T. X., & Vandelanotte, C. (2021). Applying machine learning to identify anti-vaccination tweets during the covid-19 pandemic. *International Journal of Environmental Research and Public Health*, *18*(8). https://doi.org/10.3390/ijerph18084069

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, 30*.

Wilson, S. L., & Wiysonge, C. (2020). Social media and vaccine hesitancy. *BMJ Global Health*, *5*(10), e004206.

Yoo, I., Bi, J., Hu, X., National Science Foundation (U.S.), & Institute of Electrical and Electronics Engineers. (2019). Predictive modelling of stigmatized behaviour invaccination discussions on Facebook. *Proceedings, 2019 IEEE International Conference on Bioinformatics and Biomedicine : November 18-21, 2019, San Diego, CA, USA*.

Yousefinaghani, S., Dara, R., Mubareka, S., Papadopoulos, A., & Sharif, S. (2021). An analysis of COVID-19 vaccine sentiments and opinions on Twitter. *International Journal of Infectious Diseases*, *108*, 256-262.

Yuan, X., & Crooks, A. T. (2018). Examining online vaccination discussion and communities in Twitter. *ACM International Conference Proceeding Series*, 197–206. https://doi.org/10.1145/3217804.3217912

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhut- dinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *In Proceedings of the IEEE international conference on computer vision*, pages 19–27.

Zhang, L., Fan, H., Peng, C., Rao, G., & Cong, Q. (2020). Sentiment analysis methods for hpv vaccines related tweets based on transfer learning. *Healthcare (Switzerland)*, *8*(3). https://doi.org/10.3390/healthcare8030307

# PUBLICATIONS

Hossain, S. M., & Sulaiman, S. (2022). Detecting Public Outlook Towards Vaccination Using Machine Learning Approaches: A Systematic Review. In *International Conference of Reliable Information and Communication Technology* (pp. 141-150). Springer, Cham.