**A close-up of a logo

Description automatically generated**

**The application of natural language processing techniques and forecasting approaches to covid 19 data with focus on Ireland**

by

Gavin Davis

Supervisor: Dr Catherine Mulwa

Presented to CCT college Dublin in fulfilment of the degree Master in Science in Data Analytics

Submitted: September 2023

**Abstract**

**1.Topic area**

Data analytics can assist greatly in understanding trends, temporal change and opinion of many different areas of interest. The Covid19 pandemic was something which affected all the world for nearly 3 years. Specific areas of data analytics played a pivotal role in collecting and analysing relevant information to assist and combat the pandemic in an effort to alleviate the damage. A huge amount of data was collected during the pandemic on covid cases, mortality, vaccine uptake, vaccine sentiment and more. Now post pandemic peak we have access to massive datasets related to the pandemic. One problem area during the pandemic was vaccine uptake and hesitancy. Vaccines are administered to help reduce the severity and the spread of infection. Thus vaccination is very beneficial to quell the spread of infection and alleviate the pressure of the pandemic. This project aims to apply data analytical tools such as sentiment analysis and forecasting approaches to understand and predict trends in vaccination sentiment and uptake. Tweet data will be used to generate a predictive vaccine sentiment model with several models being implemented to achieve a model with the highest accuracy. Vaccine uptake will be analysed in Ireland using publicly available datasets from data.gov by using forecasting approaches such as time series analysis. The main focus of this project is sentiment analysis and forecasting approaches as applied to covid19 vaccine data, with the aim of generating useful predictive models for prediction of pandemic related metrics. Finally, the project will focus on understanding sentiment around vaccines by carrying out primary research in the form of a Vaccine awareness survey. The sentiment will be correlated with other factors collected during the primary research phase. A comparative analysis of the sentiment collected from the primary research and the sentiment from the tweet data will be used to understand Irelands position versus the rest of the world.

This project focuses on key methods of data analytics as applied to covid19 data, with primary research carried out to understand vaccine sentiment in Ireland and potential contributing factors:

1. Sentiment analysis applied to covid19 tweet data
2. Time series analysis applied to covid case numbers/vaccine uptake levels to generate a machine learning models which can predict cases/vaccinations effectively

Each aspect of the project will be addressed as outlined in the research objectives 1-3 below.

**Research question :**

Can data analytic approaches such as natural language processing and forecasting approaches be explored to generate models for prediction of covid 19 related metrics such as vaccine sentiment and levels?

**1.1 Research objective 1**

**The aim of this research objective is to investigate vaccine sentiment in relation to covid 19 harnessing sentiment analysis**.

The first research objective of this project is to investigate and evaluate the current state of vaccine sentiment post pandemic using sentiment analysis on vaccine related tweets. Tweets will be collected using the Twitter API before using natural language processing techniques to prepare and clean text data for polarity measurements using methods such as Textblob and VADER. Tweets collected will be processed into a new dataset. The likely outcome will be N understanding of the current state of opinion on vaccines. In line with this, the sentiment data will then be used to train a machine learning models to predict sentiment of new text data. Different machine learning algorithms will be applied to the sentiment data for comparison and to achieve a model with the highest accuracy. This data will then be compared with sentiment data collected from the primary research carried out in research objective 2. As a part of this research objective a comprehensive literature review will be written on the state of the art in natural language processing and sentiment analysis, as well as its application to vaccine sentiment related to covid 19.

**Deliverables:**

* New tweet dataset about covid 19 vaccines
* Understanding of current sentiment around vaccines
* Comparison of best models to use in prediction of sentiment data
* Accurate final model with the highest accuracy
* Comprehensive literature review of sentiment mining techniques, and sentiment analysis as applied to covid 19 data

**1.2 Research objective 2**

**The aim of this research objective is to use a survey as a primary research method to investigate the current sentiment in Ireland around covid 19 vaccinations**.

This part of the research will be addressed by creating a short survey to be filled out to gather responses such as whether people feel they understand vaccines more or less post pandemic, which employment category they fall into, and whether they would be likely to receive a vaccine if another pandemic were to occur. The main information of interest will be gathered in the final section which is whether people feel positive, neutral or negative towards vaccines. This data will be correlated with other metrics collected in the survey such as employment background. Many factors were highlighted as contributing to covid 19 vaccine hesitancy and thus this survey provides an opportunity to correlate sentiment with other data collected such as employment. The two main subgroups in question will be people from a science/health background and non-science/health background. The sentiment will also be compared against the sentiment analysis performed in RO1. The data generated from this questionnaire will be analysed and presented in a non-bias manner to understand whether there are differences in vaccine hesitancy and awareness between different employment subgroups as mentioned above .The projected outcome of this research objective will be a new understanding of whether employment in health/science sectors contributes to vaccine sentiment and awareness. Given vaccine hesitancy and the factors which contribute to it were an intense area of investigation during the pandemic, relevant literature will be collected and presented in the literature review on the factors which contribute to vaccine uptake and hesitancy.

**Deliverables:**

* New dataset on vaccine sentiment in Ireland related to covid 19
* Understanding of the contribution of employment background, if any, to vaccine sentiment
* Comparison of sentiment analysis on covid 19 vaccine related tweets with sentiment in Ireland generated from primary research survey

**1.3 Research objective 3**

**The aim of this research objective is to investigate whether forecasting approaches such as time series models can be applied to covid19 vaccine data to generate accurate forecasting models.**

To carry out this aspect of the research project data will be used from data.gov on vaccination levels from Ireland. The dataset acquired from gov.ie will be explored using exploratory data analysis (EDA) and cleaned for implementation into forecasting models. Visualisations will be used to interpret the data and the accuracy of any resulting models. Different time series forecasting models will be applied to the data and compared in order to achieve the most accurate model, while comparing accuracy metrics of each respective model. The likely outcome of this part of the research will be the generation an accurate predictive model which could assist the understanding and prediction of vaccine levels in Ireland during the pandemic over time. These models will be generated using publicly available data from data.gov. The literature review carried out as a part of RO1 will cover the theory and state of the art behind forecasting approaches with emphasis on time series forecasting and the models which have been applied to covid19 data.

**Deliverables:**

* Accurate forecasting model for the prediction of vaccine trends in Ireland
* Comparison of forecasting models as applied to covid 19 data in Ireland
* Comprehensive literature review of forecasting models used in covid 19

**1.4 Relevance of project and target stakeholders**

The relevance and applicability of this research project will be an understanding of data analytical techniques which can be applied to covid19 data in order to drive decision and policy making for any future relapse or alternative pandemics. The project will use sentiment analysis and primary research to understand the current state of opinion and sentiment towards vaccines in the hope that any models and insight generated can be effectively harnessed to make decisions regarding health measures by science/health sector professionals. The application of forecasting approaches such as time series forecasting to highlight models with accurate predictive power for covid 19 vaccine levels will showcase which models are best for future use in the prediction of such metrics.

The most likely stakeholders with weighted interest in the outcome of this project would be the healthcare industry, and the benefit these models could bring to the sector. On a broader scale the pandemic effected all the world, thus developing new and improved methods for predicting future trends to insight proper planning will be of great benefit to the general population.

**2.Methodology**

**2.1 Project management framework**

The CRISP-DM approach is widely used as the industry standard set of criteria for carrying out a data project. It is the most widely used and complete methodology to carrying out a data mining project in comparison to SEMMA or KDD (Schröer, Kruse and Gómez, 2021). Thus the CRISP-DM approach was chosen as the guiding framework for this research project and its various phases and research objectives as outlined in the project outline below.

**Diagram

Description automatically generated**

Figure 2.1: A schematic representation of the approach taken for this study using the CRISP-DM methodology**.** 1-Business understanding describes getting an overview of the business and/or organisation in question, describing the projected project goals and expected outcome. 2-Data understanding describes collecting the data, exploring it and describing it using statistical analysis and visualisation. 3- Data preparation describes preparing the data for modelling by cleaning and feature engineering making the available data suitable to be used in ML models. 4-Modeling describes choosing an appropriate model that that fits the initial question and the gained understanding of the data explaining the choice and the parameters set. 5-Evaluation describes evaluating the results and discussing them in line with the objectives of the projected outcome. 6-Deployment presented as a final report or software component, including the plan for deployment and how to monitor and maintain (Schröer, Kruse and Gómez, 2021).

**2.2 Project outline**

****

Figure 2.2: Outline of project

**2.3 Datasets**

The tweet dataset used for sentiment mining generated as a deliverable of this project by collecting tweet data using the Twitter API basic access account paid for with personal funds. This allowed the mining of up to 10,000 tweets per month at a cost of 100 USD. The basic access only allowed access to recent tweets search which allowed collection of tweets from 7 days previous from the time of collection. The primary research generated a dataset used to compare sentiment in Ireland. Finally, the vaccination percentage dataset used to generate forecasting models was acquired from data.gov.

**2.4 Programming tools**

For statistics, data preparation/visualisation, and machine learning anaconda navigator was used along with Jupyter notebook as a coding interface. The language of choice for the project was python. All code files and outputs are provided alongside this report.

**2.5 Version Control**

In order to maintain work completed and store completed tasks necessary for the project a GitHub repository was set up. Commits where made on an iterative basis to save the progress of each step of the project so not to lose any progress made. GitHub link:

https://github.com/GavnDavisCCT/Masters\_project\_Covid19

**2.6 Collection of Tweet data using Twitter API and preprocessing**

**2.7 Machine learning and deep learning models**

A number of machine learning models were applied to the different datasets used in this project. For sentiment classification Random Forest Classifier, Multinomial Naïve Bayes Classifier, Support Vector Machine Classifier, Decision Tree Classifier, K-Nearest Neighbour Classifier, XG Boost Classifier, Extra Trees Classifier and LSTM models were used to predict sentiment.

**2.7.1 Decision Tree and Random Forest Classifier**

Random Forest Classifier is based on the Decision Tree model but instead of creating a single tree with multiple splits it creates a number of trees which operate as an ensemble, meaning the creation of more models to generate more accurate results (*Random Forest Algorithm for Machine Learning | by Madison Schott | Capital One Tech | Medium*, 2019). Both are supervised machine learning techniques. Each Tree predicts the class being predicted and the majority predicted class becomes he models prediction for that class. Like Decision Tree Classifier, Random Forest Classifier uses the Gini index to decide how each node branches in each of the decision trees shown in Eq.1 (*Random Forest Algorithm for Machine Learning | by Madison Schott | Capital One Tech | Medium*, 2019).

A mathematical equation with numbers and symbols

Description automatically generated

Equation 1: Gini index formula for calculating splits in decision tree classifier and multiple trees in Random Forest Classifier. This formula uses the class and probability to determine which of the branches is more likely to occur .

A diagram of a tree

Description automatically generated

Figure 2.3: Schematic of random forest architecture and prediction method (*Random Forest Algorithm for Absolute Beginners in Data Science*, 2022)

**2.7.2 Naïve Bayes Classifier**

Naive bayes classifiers are based on Bayes Theorem assuming each point of data is completely independent taking into account whether probability of an event is True to make predictions about whether new data points are True. Multinomial Naïve Bayes is often used for Text Classification tasks. It also relies on bayes theorem which calculates the probability of an event B occurring given A has already occurred, as shown in Equation 2 (*Naives Bayes Classifiers for Machine Learning | by Madison Schott | Capital One Tech | Medium*, 2023).

A math equation with a circle and a circle with a circle in the middle

Description automatically generated with medium confidence

Equation 2: Naïve bayes probability calculation formula (*Naives Bayes Classifiers for Machine Learning | by Madison Schott | Capital One Tech | Medium*, 2023).

**2.7.3 Support Vector Machine**

Support Vector Machine or SVM is another excellent algorithm for classification tasks such as sentiment analysis. The algorithm generates a hyperplane in between data points in a dimensional space related to the number of features. The aim of algorithm is to generate a hyperplane with the largest margin meaning the largest threshold possible between similar groups of data points as shown in Figure 2.4 (*Support Vector Machine — Introduction to Machine Learning Algorithms | by Rohith Gandhi | Towards Data Science*, 2018).

A diagram of a graph

Description automatically generated with medium confidence

Figure 2.4: Data segregation in N-dimensional space to create a hyperplane using the SVM algorithm (*Support Vector Machine — Introduction to Machine Learning Algorithms | by Rohith Gandhi | Towards Data Science*, 2018).

**2.7.4 K-Nearest Neighbour Classifier**

K-Nearest Neighbour classifier is a simple supervised machine learning algorithm which classifies data based on proximity to other data points (Müller and Guido, 2016). The most important value to choose in KNN classifier is K the number of nearest data points or neighbours used to classify a point (*Machine Learning Basics with the K-Nearest Neighbors Algorithm | by Onel Harrison | Towards Data Science*, 2018). The KNN algorithm uses different equations for the measurement of distance between data points, the most commonly used is Euclidean distance which measures the straight line distance between two points in Euclidean space. Other measures include Manhattan and Minkowski (*Most Popular Distance Metrics Used in KNN and When to Use Them - KDnuggets*, 2023).

A diagram of a diagram of a data point

Description automatically generated with medium confidence

Figure 2.5: Classification of data point using K-Nearest Neighbour algorithm (*Machine Learning Basics with the K-Nearest Neighbors Algorithm | by Onel Harrison | Towards Data Science*, 2021).

**2.7.6 XGBoost Classifier**

XGBoost has gained massive popularity in the last number of years for classification tasks due to high precision, working on a similar premise to random forests, XGBoost is a form of ensemble learning that uses multiple models, also known as base learners (similar to Trees in random forest) to predict (*Beginner’s Guide to XGBoost for Classification Problems | Towards Data Science*, 2021). The Trees in XGBoost differ from those in random forest and are called CART Trees (Classification and regression trees). XGBoost harnesses by altering decision making based on decisions made with other trees and using gradient descent to isolate and remove trees which are poorly performing trees are isolated and removed. Through incorporating regularization XGBoost avoids overfitting, improving overall model performance, shown by consistently outperforming other machine learning models in predictive tasks. The speed at which the XGBoost algorithm work can be attributed to parrallelization which enables the boosting process to work in parrellel thus reducing the training time for the model (*XGBoost: Everything You Need to Know*, 2023).

A diagram of a data processing process

Description automatically generated

Figure 2.6: Schematic of XGBoost algorithm (*XGBoost: Everything You Need to Know*, 2023.

**2.8 Sampling strategy**

The proposed sampling strategy for the primary research carried out in this project will be **non-probability sampling**. The entire population of Ireland was affected by the covid19 pandemic and therefore taking a sample from the population using the **convenience sampling type** should yield reflective data. The questionnaire will be sent to people which are close to hand that work in a number of different employment backgrounds. Effort will be made to have the questionnaire completed by people working in the science/health area to satisfy whether there is a difference between people’s views (hesitancy, uptake and awareness) who come from a science/health background and a non-science/health background. Given the nature of RO3, **quota sampling** will also be incorporated into this sampling process, given a number will be selected from both scientific/health and non-scientific/health backgrounds to get survey responses. It is the aim of the primary research to achieve at least 40 responses to the survey. Taken the aim of 40 responses to the questionnaire, 20 responses from each subgroup, science/health and non-science/health will be gathered. The questionnaire will be sent out to a population of people which are close to hand in order to get the responses back as soon as possible for analysis and incorporation into the project report.

**2.9 Primary research methodology**

The primary research will be performed using a carefully designed and simple survey. The preliminary considerations for the survey design will be what information will be gathered that will help achieve RO3 of the project, which is to evaluate the current vaccine sentiment of a cohort of people in Ireland. Extending this it is the aim to understand whether distinct subgroups understand whether there is a difference in sentiment and awareness within people from different employment backgrounds. The target audience will be people who received the covid19 vaccine or did not, and will be chosen in a non-discriminatory fashion except for one caveat, which is, an attempt will be made to have the survey completed by people from both non-scientific/health and scientific/health employment backgrounds to assess whether there are differences in their views on vaccines and/or the sentiment around vaccines. Given the type of information being asked could be considered sensitive, people will be asked if they are comfortable answering questions on the topic of covid19 vaccine prior to administering the questionnaire. Given the potential sensitivity of the topic the questionnaire will be anonymous. The questionnaire will be developed on an online platform and sent out via email/message which could be completed on a computer or phone. The design will be simple but professional, with the length being kept to a minimum while also obtaining the level of information required by the study. The title will be carefully chosen so not to dissuade or lead any potential responders.

Each question will be designed to add value to the information gathered to be used in the analysis of the data, comparisons and correlations can be made through gathering answers to different questions, such as, correlating age with hesitancy (Question 2 with 6). No person which the questionnaire will be administered to will be unable to answer any question. The response to the questions will be closed-ended and multiple choice so to stream line responses and gather consistent usable data. Some responses may be dichotomous, either yes or no. The closed ended response make qualitative and quantitative analysis easier, and reduce the introduction of any error, making all responses between subjects comparable. The questions will be designed to be simple, non-leading, non-loaded and understandable for a wide audience so to maximize the level of accurate responses. The sequence of questions will be considered so to guide the subject through the questionnaire in the most comfortable way possible. The most sensitive and important questions will be placed near the end and the subject will be thanked for their responses. After the questionnaire is fully designed it will be pretested before actually carrying out the study to identify any potential unseen issues and amend them.

It is the aim of this primary research section to gather at least 40 responses to the questionnaire. The questionnaire will not be sent out cold and will be administered to a populations which is close to hand, as outlined in the sampling methodology.

**Survey questions:**

1. What age bracket do you fall under? (discussion point in findings about age and vaccine hesitancy and awareness)
2. Which country were you born?   (correlating with other answers, easy lead in question)
3. How many people, including yourself, live at home with you? (discussion point in findings about whether people who live in larger homes were more or less hesitant)
4. Which option ***provided*** does your profession fall under? (Broad range of titles supplied, other option given with a follow up question asking to describe profession)
5. If you answered other to the last question, could you describe the area you work in?
6. Have you received a covid19 vaccine? (yes/no/rather not say)
7. Were you hesitant to receive a covid19 vaccination? (yes/no/rather not say)
8. Did you understand how vaccines worked prior to the pandemic? (Yes/no)
9. How would you describe your understanding of vaccines post pandemic compared to pre pandemic? (Very good, good, okay, poor, very poor)
10. Would you like to understand how vaccines work so that you can understand the effect they may have on you? (yes, no, impartial)
11. In your opinion, was the level of public awareness raised around vaccines was sufficient during the pandemic?(Will give 5 options, very good, good, okay, poor, very poor)
12. Were another pandemic occur, would you be likely to receive a vaccine were another pandemic to occur? (yes/no/rather not say)
13. Which option best describes your feelings towards vaccines? (very positive, positive, neutral, negative, very negative)

A screenshot of a test

Description automatically generated

Figure 2.3: Introduction to the survey sent out via email to participants

**3.Literature review**

**3.1 Background to problem area**

The Covid19 pandemic affected the entire world, bringing economies, trade, employment, education, and health a ream of issues that have not been experienced before. To date, at the time of writing this, 683, 644, 472 coronavirus cases have been recorded with a mortality of 6, 829, 253, approximately 1%. The Covid19 pandemic is on-going but has reached levels in most countries where some level of normalcy has returned. For many, it was the focus of daily life for 2 years, constantly exposed to case levels, mortality and the situation in other countries on every media outlet available. Given a pandemic of this magnitude has never occurred before, and the ability to gather and process data is more efficient than ever, the pandemic shone a light on data analytics, and its ability to digest, and purpose data to predict the spread, levels of vaccine uptake, sentiment and future waves of the virus. Although infectious diseases such as SARs, H1N1, Ebola and Zhika viruses have threatened this level before, with a prospective worse outcome given their increased severity compared to covid19, nothing of this magnitude has been seen or recorded before. Thus covid19 is now the model pandemic which will assist the domain of data analytics in generating valuable insights and predictive models to understand the worlds response, failings, areas of concern, vaccine uptake, levels of infrastructure required to combat such levels of widespread infection, to name a few. Hence, the focus of this piece will be how data analytics is being used to study covid19 and generate valuable insights and tools to combat the ongoing but subdued pandemic that brought the world to a standstill, highlighting areas relevant to this project. Key areas within the domain of data analytics which this review will focus on are natural language processing, specifically sentiment analysis, and forecasting approaches such as time series forecasting as applied to covid19 data and related data.

Healthcare is slowly but surely making the transition to data driven decisions for nearly every facet of the sector, whether it be for patient diagnostics, drug/small molecule discovery, disease severity and onset and drug efficacy (Cascini *et al.*, 2021; Chauhan *et al.*, 2021; Savage, 2021). The covid19 pandemic brought to light the power of data analytics and its invaluable contribution to decision and policy making. Data analytics was pivotal in the response and eventual subduing of the covid19 pandemic. A number of tools were used in assisting with the collection, processing, understanding, translation and dissemination of how the pandemic was unfolding, through assessing parameters such as case numbers, severity, vaccine uptake and sentiment. Now with the height of the pandemic seemingly behind us, data generated during the pandemic can be a harnessed to generate deep insights and predictive models that may assist in the future. In line with the aims of this project, sentiment analysis and forecasting approaches will be harnessed, applied and explored using covid 19 data metrics, and thus this review will focus on the state of the art in these techniques, and their application to covid 19.

**3.2 Natural language processing**

Natural language processing is a sub field of AI and linguistics, stemming from the 1950s, which offers a way for a computer to interpret natural languages (Khurana *et al.*, 2023). Given the idiosyncratic nature of languages such as English and common phrases, NLP offers a way in which to vectorise words and phrases making it digestible for algorithms, such as machine and deep learning models. Before delving into the advanced applications of sentiment analysis, it is important to first understand the fundamentals on which it operates, which is NLP. As described in the quote below from Dipanjan Sarkar’s book on Text analytics from 2016, Natural language follows rules and syntax which can be exploited through commonalities and converted into data which can be analysed mathematically (Sarkar, 2016).

“Textual data is unstructured data but it usually belongs to a specific language following specific syntax and semantics. Any piece of text data—a simple word, sentence, or document—relates back to some natural language most of the time” -Dipanjan Sarkar ,2016

There are several applications for natural language processing including predictive text, email filtering, document analysis, social media monitoring, chatbots, language translation and sentiment analysis (ref). More recently, a massive leap forward was made in using predictive text and NLP for the creation of Large language models such as Chat GPT-4 by Open AI. This technology harnesses several key processes of natural language processing to allow the user to input information and receive relatively accurate output for various queries or prompts. Natural language processing was applied extensively during the covid 19 pandemic to assist in digesting massive amounts of text data from sources such as social media, electronic health records, scientific investigation literature and health agency guidelines (Al-Garadi, Yang and Sarker, 2022). The various applications of NLP in covid 19 are shown in Figure 3.1.

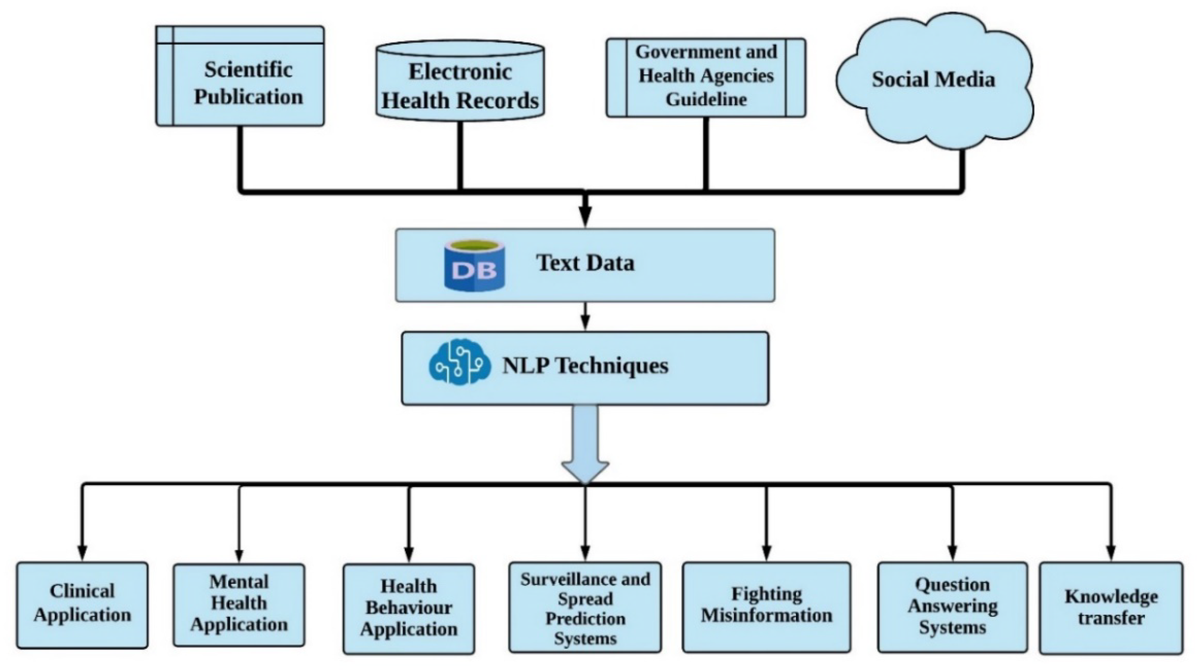


Figure 3.1: Showing summary of NLP methods applied during covid 19 (Al-Garadi, Yang and Sarker, 2022)

A key area which is often used in NLP is sentiment analysis (Gonçalves *et al.*, 2013). Several excellent machine learning and deep learning approaches have been used to develop pre trained models to carry out natural language processing tasks such sentiment analysis like word2vec, GloVe, BERT (Koroteev, 2021). TextBlob and VADAR, are lexicon based NLP methods,, non- machine learning model based methods, rely on a large dictionary of words which the algorithms already possess, and which the training data provided is compared against to perform NLP tasks (Fang and Zhan, 2015; Kolchyna *et al.*, 2015; Rodríguez-Ibánez *et al.*, 2023) . Given an objective of this project is to mine sentiment from a new twitter dataset generated with the twitter API based on current covid 19 vaccination tweets, these techniques will be explored and refined for their potential to generate the most accurate model possible for classification of sentiment of the problem area dataset.

**3.3 Sentiment analysis**

Sentiment analysis is a key area of natural language processing which enables the mining of opinion and feelings around a certain topic. This enables more efficient decision making when it comes to policy and business decisions. Sentiment analysis was applied extensively during the pandemic to mine public opinion particularly in relation to vaccines, this will be discussed in a later section. Simply put, sentiment analysis relies on the ability to process textual data to its rudiments and convert the text data in to numeric type data which can be digested and quantified by various algorithms, such as the ones mentioned in Section 3.2 (Gonçalves *et al.*, 2013). Sentiment analysis is an area of natural language processing which can determine the positivity, negativity and neutrality of textual data to understand the overriding feeling and opinion of a certain topic (*A Comprehensive Overview of Sentiment Analysis*, 2022). The areas which sentiment analysis can be applied to are endless, including but not limited to, consumer reviews, product reviews, brand reviews and public opinion on social media websites like Twitter and Facebook (Gonçalves *et al.*, 2013; Fang and Zhan, 2015; Cepeda and Jaiswal, 2022; Rodríguez-Ibánez *et al.*, 2023).

The pipeline for performing sentiment analysis involves a number of key NLP tasks and is generally similar across different methods, with a few alterations depending on the model being applied. A critical step common to most sentiment analysis approaches is the cleaning and preparation of the data. A typical pipeline for cleaning and preparing data for sentiment analysis is shown below:

A diagram of a diagram

Description automatically generated with medium confidence

Figure 2: Steps cleaning and processing text data for sentiment analysis (*A Sentiment Analysis Approach to Predicting Stock Returns | by Tom Yuz | Medium*, 2023)

The processes required to clean and prepare text data for sentiment analysis are enabled by a number of python libraries, the most prominent of which being Natural language processing toolkit, abbreviated as NLTK (Wang and Hu, 2021). Sentiment analysis involves processing text data by removing noise from the data which takes the form of stop words, such as ‘and’ which don’t add much value in terms of identification to a body of text, thus they are removed. Another technique used is stemming and lemmatisation which reduces similar words to a common rudiment in order to group them together for comparative purposes again reducing the noise, taking the words playing, play and playful, and combining them into a single representative word of play . Regular expressions are employed which are common sequences of characters found in natural language which can again remove noise, removing special characters like hashtags or converting all letters to lower case for comparative purposes (*Regular Expressions — An excellent tool for text analysis or NLP | by Niwratti Kasture | Analytics Vidhya | Medium*, 2019). Following this Tolkenization, the process of splitting a body of text into individual components is performed, followed by vectorising the words or phrases so they may be applied to models. Finally, and most importantly, tools such as TextBlob and VADER, a python library can be used to gauge the polarity, which is a rating between -1 and 1, describing the relative positivity, neutrality or negativity of the text. This rating can be highly useful when it comes to mining opinion around a certain topic as, for instance, the number of positive, negative and neutral tweets can be assessed around vaccinations to understand public opinion on the topic (*TextBlob | Making Natural Language Processing easy with TextBlob*, 2022).

As previously mentioned, there are two methods used for sentiment analysis: Lexicon-based and machine learning based. Lexicon-based uses a dictionary of words that are compared against the text being analysed enabling the assignment of polarity measurements associated with positive, negative or neutral words (Kolchyna *et al.*, 2015). This method which is used by TextBlob and VADAR, is commonly used The second method is machine learning based, with examples being methods such as BERT. BERT uses a multi-layer bidirectional transformer encoder which represents the input text in high dimensional space, enabling understanding of the entire context of each word in a collection of text such as a tweet (*Text Classification with BERT. In this tutorial, we will use BERT to… | by Khang Pham | Medium*, 2023) When the model is trained, new text can be given to the model and the model can make predictions based on its training. Many pre trained models exist for machine learning based approaches such as BERT. The two techniques can be used together, and often are, whereby text data is assigned a sentiment using Lexicon-based approaches and then a model is trained with this data to enable predictions of new data. Within this project it is the aim to do this using covid19 vaccine related tweets and to generate a predictive machine learning model which could be used to assign sentiment to new data of the same topic.

Many different machine learning models have been applied to perform sentiment analysis. Given that sentiment is a classification task for understanding the relative positivity or negativity of textual data many classification models such as Random Forest Classifier (RF Classifier) (Iwendi *et al.*, 2020), Decision Tree Classifier (DTClassifier) (Chinnasamy *et al.*, 2022), Support Vector Machine (SVM) (Binti, Nasir and Palanichamy, 2022), Naïve Bayes Classifier (NBClassifier) (Iksan *et al.*, 2021), XGboost Classifier (XGBClassifier) (Huang and Wang, 2023) . As well as these machine learning models, deep learning techniques have been applied to sentiment analysis such as the popular Recurrent neural network (RNN) model known as Long short term memory or LSTM and the relatively new BERT model. These models have shown excellent accuracies for classifying sentiment in the literature, and importantly for this project, have been applied to mine sentiment for covid 19 related text data. As well as this combinations of sentiment mining techniques, vectorization techniques and different machine learning models have been applied to achieve optimal accuracy metrics (Qorib *et al.*, 2023)

**3.4 Sentiment analysis and covid 19**

Gaining an insight into public sentiment surrounding the pandemic and vaccines was imperative to gain insights and understand how vaccines would be taken up by the general population. Several interesting studies have arisen from sentiment analyses performed on social media, news and more in relation to the covid19 pandemic, vaccinations in particular. Vaccinations were a topic of constant discussion throughout the course of the pandemic, dividing opinion amongst the general population.

In one particularly interesting study the sentiment of covid19 was compared between twitter and reddit, two popular social media outlets (Melton *et al.*, 2022). In this study, 9.5 million tweets and 70 thousand reddit comments were compared between January 1st 2020 and March 1st 2022. Both platforms shared similar sentiments with Twitter being more negative (54.8% positive) than reddit (62.3% positive), with the sentiments of both platforms displaying similar behaviour related to vaccines during the pandemic (Melton *et al.*, 2022). Another such study analysed sentiment of Covid19 news headlines during the pandemic, harnessing the data for the generation of a highly accurate neural network model (accuracy=0.931) providing an indication of how news outlets such as CNN covered the pandemic and vaccination (Ahmad *et al.*, 2022). Overall these two studies highlight two sentiment analysis was a critical element of DA which contributed to understanding of how the pandemic was unfolding.

A number of models have been applied to mine covid 19 sentiment and related topics. For sentiment analysis, it is common practice to apply classification models such as Naïve bayes Classifier. A study in 2021 by Iksan *et al.* mined sentiment from a public tweet dataset based on covid 19. Their experimental results showed the model accuracy was 0.86 or 86%, with a precision of 0.827, recall of 0.687, and an F-score of 0.749 (Iksan *et al.*, 2021). Another study by Ressan *et al.* in 2022 used multinomial Naïve bayes classifier on a covid 19 tweet dataset consisting of 10,000 tweets. Their model achieved an high accuracy of 91.6% which highlights the power of the multinomial naïve bayes classifier in sentiment classification tasks. The authors site that enhanced accuracy could be reached with a larger dataset (Ressan and Hassan, 2022). Taken together these results show promise in using naïve bayes classifier and multinomial classifier for sentiment prediction, and specifically in relation to covid 19.

Other models which have been applied to mine sentiment on covid19 based tweets are random forest, logistic regression, decision tree, linear SVC and multinomial naïve bayes. In a paper published in 2023, Qorib *et al.* performed a comparative analysis of the aforementioned models along with sentiment classification techniques such as Azure machine learning, VADER and Textblob, while also varying the vectorization methods Doc2Vec, CountVectorizer and TF-IDF (Qorib *et al.*, 2023). They acquired data daily from September 26, 2021 to November 7, 2021, combining the collected tweets into a dataset of 42,790 tweets. The best model post stemming consisted of combining TextBlob+TF-IDF+LinearSVC, with an accuracy of 0.94860, approximately 95%, with the second best being TextBlob+CountVectorizer+TF-IDF+LinearSVC (Qorib *et al.*, 2023). The lowest performing model was a combination of Azure+Doc2Vec+DecisionTree with an accuracy of 41.42% and a precision of 0.38715. Comparing model results between stemming and lemmatisation, the authors observed an increase in accuracy and precision upon lemmatisation with the TextBlob+TF-IDF+LinearSVC , with an accraucy of 0.96529 or 96.5%. Finally combining stemming and lemmatisation, furtherly increased the accuracy of the TextBlob+TF-IDF+LinearSVC from 0.96472 to 0.96752 (Qorib *et al.*, 2023). Taken together these results indicate that the combination of sentiment classification methods, vectorisation techniques and model choice can have a marked effect on the overall accuracy of sentiment classification models.

Many studies have taken advantage of the multiple sentiment analysis techniques mentioned in the previous sections. Rustam et al. 2021 give a comprehensive overview and compare methods used in covid-19 tweet sentiment modelling (Rustam *et al.*, 2021). They make use of the tweet IDs provided by the IEEE data port to scrape tweets with their own in house Twitter scraper and measure tweet polarity, applying a number machine learning methods such as RFClassifier (accuracy = 0.86), XGBoost Classifier (accuracy = 0.85), SVC (accuracy = 0.85), Extra trees Classifier (ETC) (accuracy = 0.88\*), and DTClassifier (accuracy = 0.83). They compare their own sentiment classification (SS2) vs that of the sentiment calculated in the IEEE data port repository (SS1) and conclude the performance of the models used were similar with the two sentiment classifications. Interestingly, they concatenated the results of the Bag of words (BoW) and TF-IDF, two word to vector algorithms, and saw an overall increase in the performance of each model used, for example the ETC model went from an accuracy of 0.92 to 0.93, thus increasing the accuracy of the model. This is an interesting finding in that creating more dimensions in the data by vectorizing the data by two methods can be used to increase the accuracy of a model. Also within this study, LSTM is used for sentiment classification and saw an overall decrease in accuracy <0.8. Rustam *et al*. conclude the study by stating that the ETC model is of the highest accuracy with an accuracy score of 0.93 (Rustam *et al.*, 2021).

Apart from classical classification models, and combinations of sentiment calculation and vectorisation, deep learning methods have been applied to predict sentiment in covid 19 from different platforms providing textual data. Mansoor *et al.* 2020, used TwitterScraper to collect tweet datasets on different covid 19 related topics such as general corona virus tweets, online learning tweets and work from tweets. The corona virus tweets dataset was used to train an LSTM model and an ANN model along with VADER to calculate tweet polarity. The LSTM showed superior accuracy of 84.5%, while the ANN showed an accuracy of 76%, showing that these models can be efficiently implemented to perform sentiment classification tasks (Mansoor *et al.*, 2020). The authors collected tweets over an extended time period analysing sentiment over the course of several months. An interesting finding was there was a marked change in the ratio of negative to positive tweets about coronavirus, in favour of negative tweet sentiment, with a more even ratio in the later months of the study for the year 2020 (Mansoor *et al.*, 2020). This study highlights not only that deep learning techniques can be applied for sentiment classification tasks but sentiment analysis can be harnessed to evaluate opinion on a certain topic over time.

A start of the art deep learning method for sentiment analysis was introduced in 2019 by google known as BERT. One particular study carried out by Muller *et al.* in 2023 developed a BERT model trained on a large corpus of covid 19 related tweets ( 160 million ) specifically designed for use in social media evaluating the performance on 5 different classification datasets, known as COVID-Twitter-BERT (CT-BERT). The models performance was measured against BERT-LARGE, a pre-trained, English language model. The model consistently outperformed BERT-LARGE and particularly when used on covid 19 specific datasets (Müller, Salathé and Kummervold, 2023). Compelling, a small subset of tweets from 2020 sampled from the CT-BERT train dataset was used to evaluate model performance against BERT-LARGE where they observed an increase in accuracy from 0.931 to 0.949. The development of these domain specific models, such as CT-BERT for covid 19 classification tasks, will undoubtedly be a valuable tool for sentiment analysis in the future, reducing computational costs for training large models from scratch.

**3.5 Forecasting models**

Forecasting models enable the prediction of variables of interest by supplying historical data to the model in an effort to predict future values. Forecasting is used by nearly every industry in an effort to predict metrics such as sales, stock price, meteorological events, sentiment and more. The great benefit of these models is insight into the future trend of the particular variable of interest to inform decisions. Such models proved valuable during the covid 19 pandemic for instance, in predicting important metrics such as case numbers, new waves, vaccine uptake levels and economic metrics (refs). This enabled informed strategic planning in the use of preventative measures such as lockdowns, travel permissions, international travel, use of face coverings and the opening and closing of different economic sectors (*COVID-19 Forecasting and Mathematical Modeling | CDC*, 2023). Now post pandemic, using the vast quantity of available datasets it is possible to fine tune forecasting models for the prediction of pandemic metrics in order to obtain models which are best suited for prediction. Given this project aims to use forecasting models for the prediction of vaccination levels during covid 19 and compare to real data to evaluate accuracy, the best models can highlighted and compared. The next section details the forecasting approaches used during the pandemic for prediction of different pandemic related metrics. It is particularly apparent in from the literature that time series forecasting has been applied to many pandemic related metrics.

**3.6 Time series analysis**

Time series analysis describes the models applied to sequenced and temporally ordered data to generate models that can predict future events related to the event being recorded. For example time series analysis has been applied a multitude of different data types such as, meteorological data such as temperature and precipitation, as well as stock prices, and even sentiment (Hewage *et al.*, 2019). The benefit of these models is the ability to be able to predict future trends in the parameter in questions such as increases in temperature or fluctuations in stock market prices. This method of prediction is of great benefit when attempting to make informed decisions about particular metrics, informing important business decisions and policy making. There are a number of different methods for time series analysis, autoregressive moving average models such as ARIMA, and recurrent neural network based models such as Long short term memory (LSTM) models and gradient recurrent unit (GRU). Neural network approaches such as LSTMs and GRU models have gained popularity in recent years as the most accurate models when it comes to time series forecasting, however there is also strong basis for hybrid models and models which incorporate influencing factors on the metric of interest which will be discussed (Yamak, Yujian and Gadosey, 2019).

**3.7 Time series analysis in covid 19**

Given the pandemic was relatively uncharted territory for the world, as well as having the ability to collect and process vast amounts of data, another key area where DA shone and continues to produce valuable insights and models is in the generation of predictive models using machine learning (ML) and artificial intelligence (AI)(Syeda *et al.*, 2021). In a systematic review by Syeda *et al.* 2021, they highlight key areas where ML and AI were being used to combat the pandemic. Of 419 articles published between December 2019 and June 2020, three key themes arose, they were computational epidemiology, early detection and diagnosis and disease progression (Syeda *et al.*, 2021). The use of neural networks for generating advanced and highly accurate machine learning models operates on the premise of having good training data. Given the peak of the pandemic is seemingly over it has opened the opportunity for retrospective studies harnessing the large datasets generated during the pandemic for the generation of valuable models. A popular method for time series forecasting is known as Long short term memory (LSTM), a form of recurrent neural network (RNN) which can assist in predicting spatiotemporal data such as daily cases or vaccination rates (Yamak, Yujian and Gadosey, 2019) Other methods such as ARIMA and GRU may prove more accurate in generating predictive models for vaccinations and case levels during pandemics in comparison to LSTM, this will need to be verified (Yamak, Yujian and Gadosey, 2019).One such study used RNN-LSTM for covid19 infection forecasting published in January 2022 in different Indian states (Chandra, Jain and Chauhan, 2022). Shahid *et al.* in 2020 compared models which were having the highest accuracy levels for the prediction of confirmed covid19 cases. The study compared  Bi-LSTM, LSTM, GRU, SVR and ARIMA with BI-LSTM showing the highest accuracy in the prediction of deaths in China with MAE= 0.007 and RMSE = 0.007 (Shahid, Zameer and Muneeb, 2020). The study concluded that Bi-LSTM could be exploited for use pandemic metric prediction leading to better planning and management. Such models proved useful in predicting whether another wave of infection would occur informing decisions on measures to be implemented to quell the spread of disease. Now with more complete datasets available these models can be effectively probed to understand their relative accuracies and efficiencies in predicting pandemic related metrics.

ARIMA models have been used to predict the number of cases, deaths and recoveries during covid19 based off daily figures reported (Gecili, Ziady and Szczesniak, 2021; Abonazel and Darwish, 2022). An interesting study used ARIMA to model the number of covid19 cases admitted to hospital in Thailand’s first university field hospital. The model achieved accuracy metrics of *R*2 = 0.5695, RMSE = 29.7605 and MAE = 27.5102 (Somyanonthanakul *et al.*, 2022). In this study the authors enhanced the accuracy of the original ARIMA model by performing association rule mining (ARM) which led to a model with increased accuracy: *R*2 = 0.5695, RMSE = 27.7508, MAE = 23.4642. This study through the application of time series analysis and ARM discovered that patients with hospital stays longer than 14 days were healthcare worker patients and patients with underlying diseases (Somyanonthanakul *et al.*, 2022). These models highlight that time series analysis can be successfully applied to covid19 related metrics such as cases, deaths and recoveries, which highlight the usefulness of these machine learning approaches in covid19.

**3.8 Data analytics as a tool to understand drivers of vaccine hesitancy and uptake during covid 19**

An issue of concern during the pandemic was vaccine hesitancy. “Vaccine hesitancy refers to delay in acceptance or refusal of vaccination despite availability of vaccination services” (Ingram *et al.*, 2023). The key drivers of vaccine hesitancy are still being unveiled. By January 2022, 91% of the eligible population of Ireland were fully vaccinated, but it is estimated that one third of the population experienced some level vaccine hesitancy (Ingram *et al.*, 2023). As of March 2023, 72% of the world’s population are fully vaccinated against the still present covid19 virus. Many studies have been carried out to identify drivers of vaccine hesitancy in different groups. One study on a vaccine hesitant subgroup in Malaysia identified trust, perceived susceptibility to the disease and perceived benefits as contributors to vaccine hesitancy (Vaithilingam *et al.*, 2023). Another such study identified that approximately 400,000 migrants in EU or EEA countries are underimmunised for a host of infectious diseases including covid19, with barriers including language, literacy, communication, practical and legal obstacles to meet the level of safe immunisation required (Crawshaw *et al.*, 2022). Despite some resolution being brought to the pandemic, understanding vaccine hesitancy and its drivers is a critical part of designing strategies for public immunisation for potential future variants and/or new infectious diseases which threaten pandemic level spread. A key point is that, were symptoms and mortality to be more severe during the covid19 pandemic, having approximately 30% of the worlds population still unvaccinated 3 years post outbreak is unacceptable.

A number of in-depth studies have identified the key drivers of vaccine hesitancy and uptake in relation to the covid19 vaccines. Identifying the drivers of vaccine hesitancy is of great importance if another pandemic were to occur. As previously mentioned, the drivers of vaccine uptake and hesitancy are still being unveiled. One study used an online questionnaire aiming to identify the young peoples willingness to continue to receive a covid19 vaccination, also identifying things such as degree of disease anxiety which remains, vaccine brand loyalty and perceived infectability (Lee and Wu, 2023). The study identified within this group of young people women were more willing to be vaccinated than men and that of the cohort assessed they were willing to receive continuous vaccination. The idea behind the study was that young people are the most active in group activities and this should be considered when implementing effective vaccination and policy measures (Lee and Wu, 2023).

**3.9 Conclusions**

This short review has highlighted key areas which data analytics is proving invaluable in addressing and strengthening the effort to alleviate the covid19 pandemic. Firstly opinion mining using sentiment has yielded deep insights into the public views on key pandemic related topics such as vaccines. Understanding public views on this area has proved invaluable for policy making and increasing awareness. The pandemic has generated datasets which will be studied in great detail for years to come such as case numbers and vaccination rates. Developing predictive models from these datasets will prove valuable for analysing future trends in pandemic related metrics such as cases and vaccinations. Time series analysis has been used previously for modelling case numbers and vaccination rates in order to predict future trends. Now that the pandemic has dimmed and we now possess more complete datasets we may be able to achieve models with higher accuracy that may be useful in the future. Finally, a key area of study during and now after the pandemic are contributors to the vaccine hesitancy and uptake. Key drivers of vaccine hesitancy and uptake are still being uncovered. It is imperative and therefore an aim of this project to attempt to understand other areas which contribute to vaccine hesitancy, uptake and awareness. Thus, in conclusion, developing ways to both predict the spread and other pandemic associated parameters such as spatiotemporal vaccination levels for instance, alongside understanding public opinion and the key drivers of vaccine hesitancy and uptake, will prove invaluable in bringing about the most efficient trajectory out of, and in avoidance of pandemic level disease spread occurring again.

**4.Results (2000-2500 words)**

**4.1 Sentiment analysis of covid 19 tweets**

**4.1.1 Collection and pre-processing of Tweet data**

Tweepy was used to extract tweets from Twitter on covid 19 vaccines. The query was “ covid-19 vaccine”. As previously described in the methodology section, the Twitter API basic access only allowed for access to 7 days previous , using the ‘search\_recent\_tweets” method. The Tweepy paginator was using to gather more than the max results parameter limit of 100, setting the limit to 10,000. This enabled the collection of 5,620 tweets. Before preparing and cleaning the tweet data for sentiment analysis, the dataset was explored to remove null values or duplicated tweets. The dataset was reduced from 5,620 to 5,515 after removal of duplicates present. The tweet data was prepared for sentiment analysis by performing a number of operations to clean the textual data. These included removing converting to lowercase, removing special characters such as hashtags, removing stop words, removing URLs, converting multiple spaces into single spaces. The removal of stop words was enabled by NLTK library. After this Porterstemmer was used to stem the respective words into common rudiments thus reducing unwanted noise in the data.

A screenshot of a social media post

Description automatically generated

A screenshot of a medical checklist

Description automatically generated

Figure 4.1: Unclean vs clean tweet data after pre-processing with NLP techniques

**4.1.2 Calculating tweet polarity using TextBlob and VADER**

After the necessary text preprocessing steps, TextBlob and VADER were used to assess the polarity of the cleaned textual data. Both are lexicon based methods however VADER is said to be more accurate for calculating polarity of social media content (*TextBlob vs. VADER for Sentiment Analysis Using Python | by Amy @GrabNGoInfo | Towards AI*, 2022). As shown in Figure 4.2, the TextBlob and VADER polarity measurements differ greatly particularly in the magnitude of positive and negative polarity. VADER appears in most cases to accentuate the magnitude of positive and negative polarity as compared with TextBlob, an interesting difference, which could be connected with the enhanced ability of VADER to calculate polarity of social media data.

**A graph of a graph showing a number of text

Description automatically generated with medium confidence**

Figure 4.2: Comparison of polarity measurement of covid 19 tweets using TextBlob and VADER.

**4.1.3 Assessing the weight of positive, negative and neutral tweets in the dataset**

After pre-processing the tweet data and calculating the polarity of the cleaned tweet data using TextBlob and VADER, an in depth analysis to understand the distribution of positive negative and neutral sentiment amongst the tweets was performed. To do this a number of visualisation were generated using plotly and WordCloud. Figure 4.3 and 4.4 show a bar graph and pie chart which highlights the number and percentage of positive, negative and neutral tweets as calculated by TextBlob and VADER. As shown in Figure 4.3, the overriding sentiment calculated by TextBlob is neutral (45.3%), followed by positive (34.6%), with negative (20.1) being the lowest sentiment class. In sharp contrast Figure 4.4 shows the distributions of sentiment as calculated by VADER with negative (48.5%) being the overriding sentiment, followed by positive (26.3%), followed by neutral (25.2%). This is an interesting distinction, and shows that TextBlob and VADER differ in their sentiment caluclations this covid 19 tweet dataset. Upon inspection of the word clouds generated for both sentiment polarity calculation methods, shown in Figure 4.5 and 4.4, it can be seen that there is a clear distinction in the words which are being identified as positive, negative and neutral between the two methods.

**A graph showing different colored squares

Description automatically generatedA pie chart with numbers and a few red and blue squares

Description automatically generated**

Figure 4.3: Distribution of Positive, negative and neutral tweets using TextBlob.

**A graph of different colored squares

Description automatically generatedA pie chart with numbers and a red circle

Description automatically generated**

Figure 4.4: Distribution of Positive, negative and neutral tweets with VADER.

A close-up of words

Description automatically generated

A close-up of words

Description automatically generated

A close-up of words

Description automatically generated

Figure 4.5: WorldCloud of Positive, negative and neutral tweets with TextBlob.

**A close-up of words

Description automatically generated**

**A close-up of words

Description automatically generated**

**A close up of words

Description automatically generated**

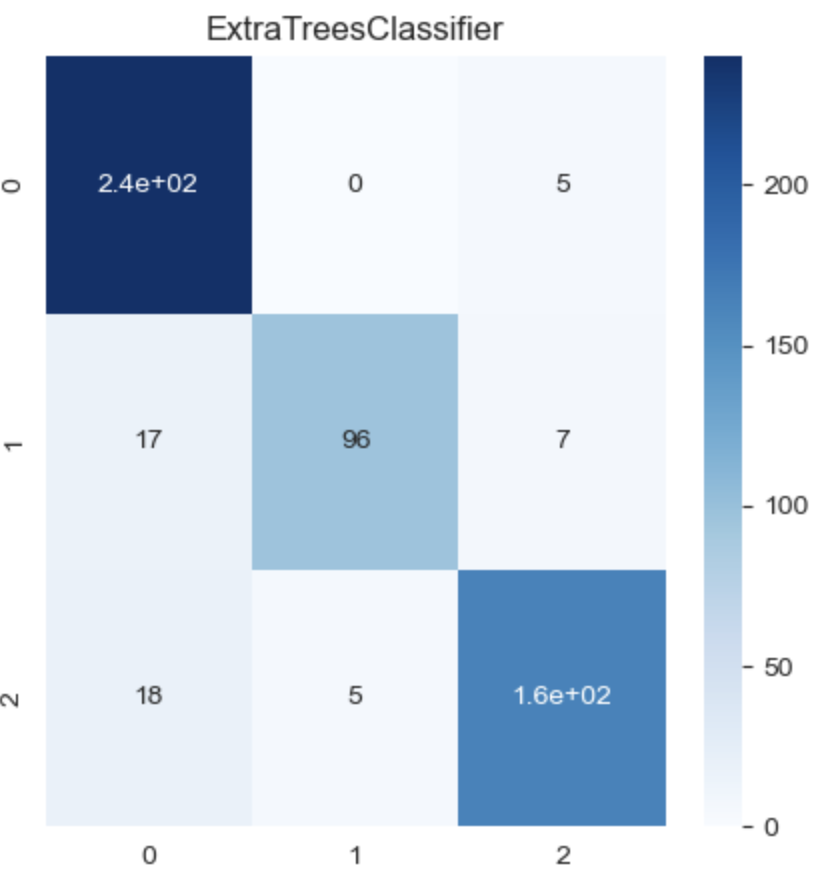
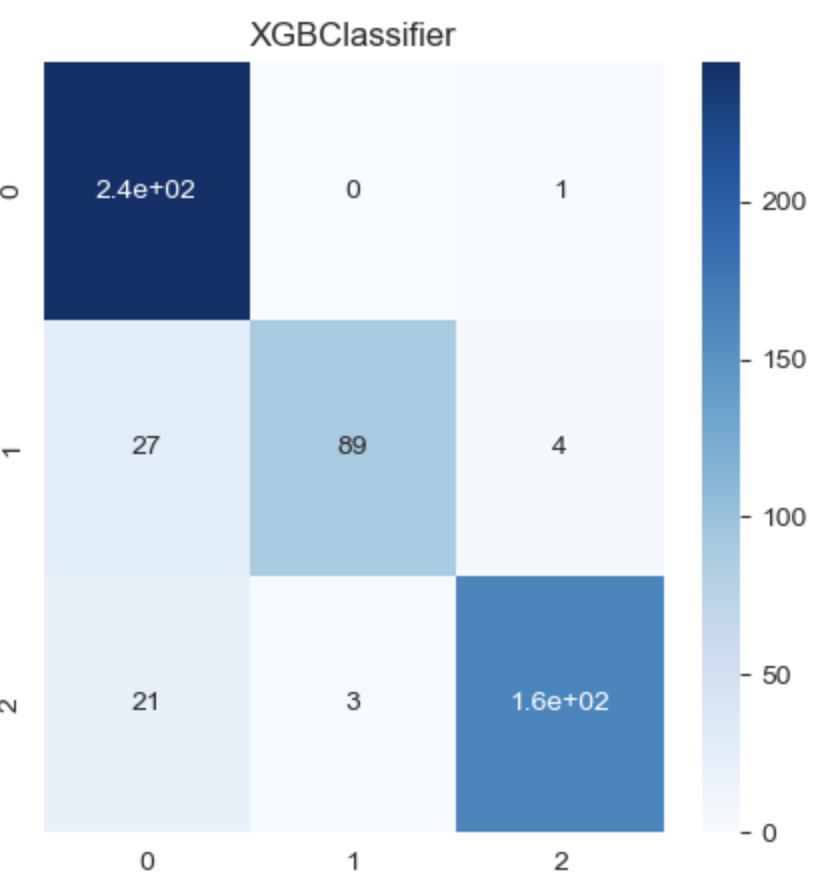
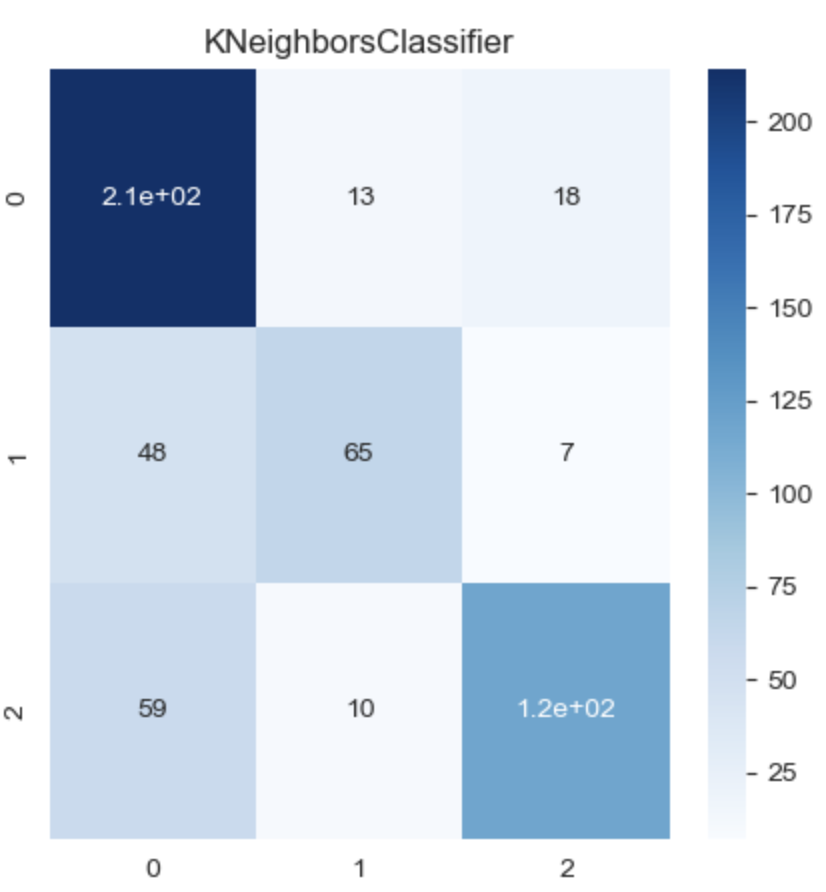
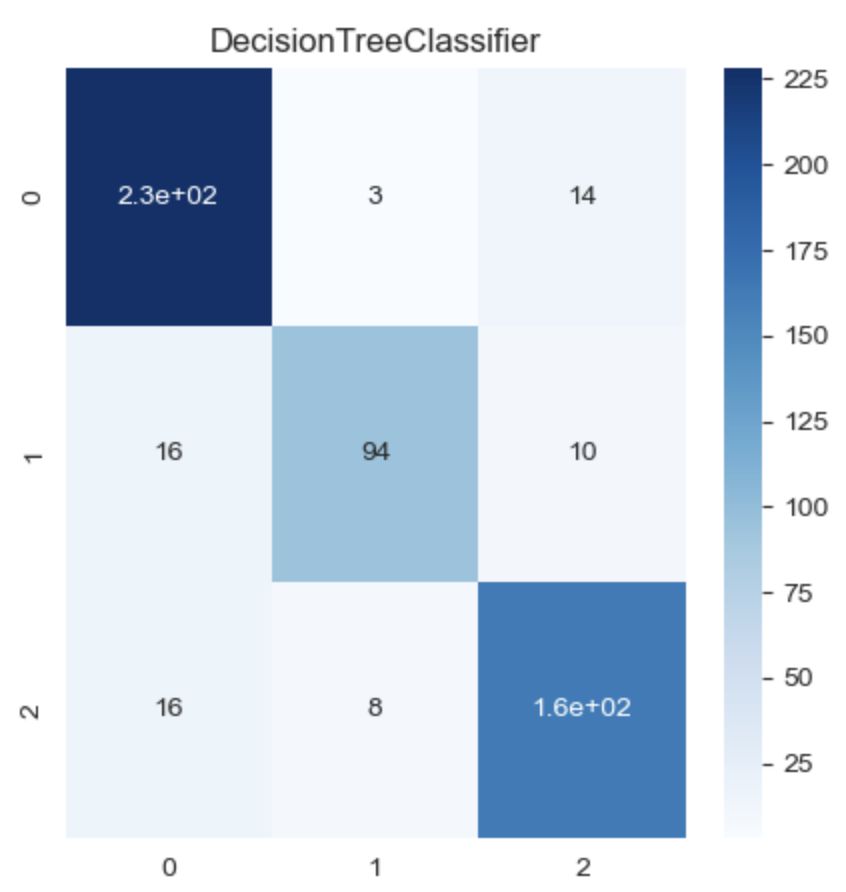
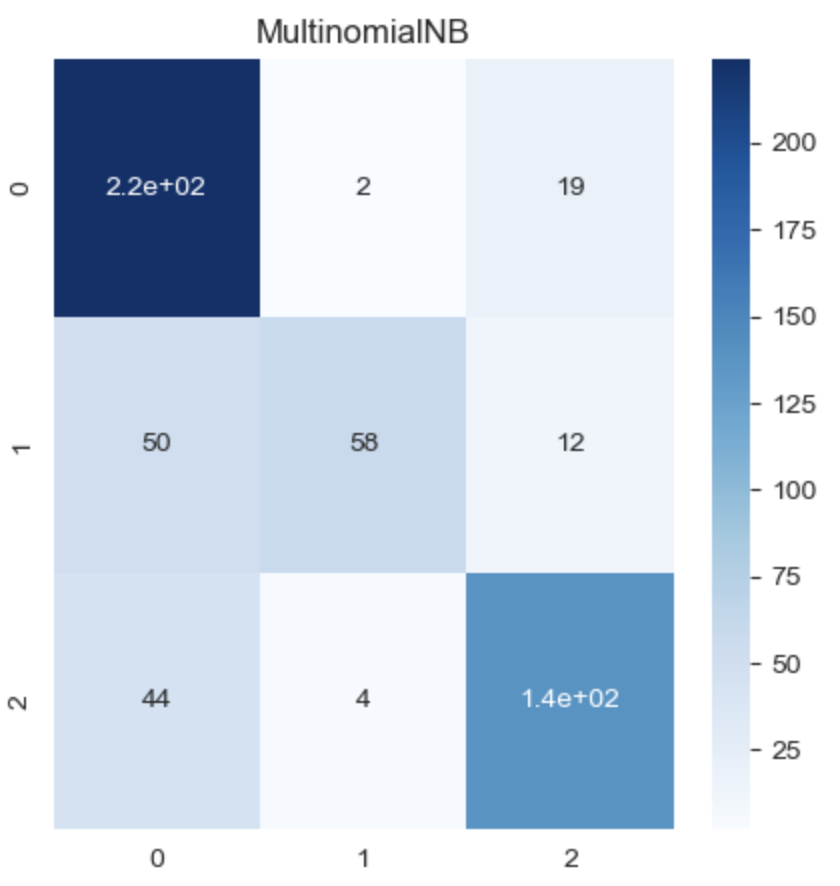
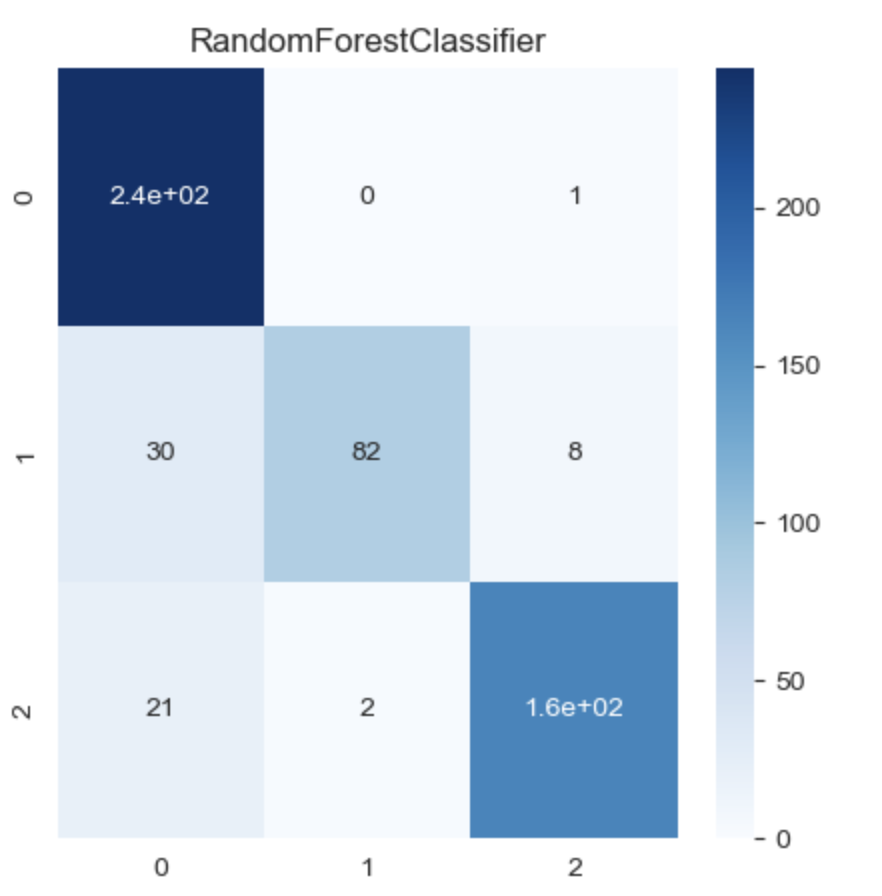
Figure 4.6: WordCloud of Positive, negative and neutral tweets with VADER.

**4.1.4 Implementation of machine learning models for tweet sentiment classification using TextBlob and VADER sentiment datasets**

After calculating he sentiment polarity scores with TextBlob and VADER and assigning positive, negative and neutral sentiment to each tweet in the dataset. Machine learning models were explored and implemented to generate accurate classification models trained on the covid 19 tweet dataset to predict sentiment. To do this, a number of machine learning models were trained using the dataset, firstly on the TextBlob sentiment classification method. As shown in section 4.1.3, the overriding sentiment in the TextBlob dataset was neutral, followed by positive, with the lowest being negative. To train the machine learning classification models the cleaned tweet data was vectorised using the TF-IDF vectorization technique from sklearn to represent works as numeric type. To implement the TF-IDF, the data was split, labelling the tweet text data as ‘X’ and the sentiment as ‘y’. The sentiment was encoded using a lambda function to assign a value of 0 to neutral, 1 to negative and 2 to positive to allow use in the machine learning model. TF-IDF was applied to the clean tweet data using a max\_features = 3000, analyzer = word, ngram\_range = (1,3) and stop\_words = English. After vectorization he data was partitioned into training and testing data using sklearns train\_test\_split, with a test size of 0.1 or 10% and a random\_state =0. A for loop was created to apply seven different classification models to the data and make predictions. The for loop included print statements to produce the relevant model metrics to assess the accuracy, precision, recal and F1-score of the models. The models applied to the data were Random Forest Classifier, Multinomial Naïve Bayes Classifier, Support Vector Machine Classifier, Decision Tree Classifier, K-Nearest Neighbour Classifier, XG Boost Classifier and Extra Trees Classifier. Classification reports were printed for each of the models and recorded as seen in Table 4.1. Confusion matrices were generated for the predictions of each model as shown in Figure 4.7. Figure 4.8 shows the summary of model accuracies obtained for each model after training and prediction with the TextBlob sentiment dataset. The model with the highest accuracy of 0.91 or 91% was the Extra Tress Classifier followed by the XG Boost Classifier.

To evaluate whether deep learning techniques such as LSTM which have been extensively applied in the literature for sentiment classification could achieve a greater accuracy than the previously implemented models, the data was prepared for implementation into an LSTM model. To use the data to train an LSTM model, further processing steps were required including tolkenizing the tweet into lists of single words, following this the length of each tweet was assessed to include a pad sequence to define the limit for all input texts, as the LSTM requires all to be the same length. The assessment of tweet length can be seen in Figure 4.9 using descriptive statistics and histograms to visualize the length of positive, negative and neutral tweets. A max word length of 50 was chosen for the LSTM implementation. Dummie variables were used to enumerate each sentiment class by using get\_dummies. The data was split as before with sklearn train\_test\_split with a test size of 10% and a random\_state = 0. TF-IDF was used to vectorize the tweet data and pad sequences were tolkenized using sklearn.tolkenizer. At this point the train, test and validation sets were generated. The LSTM architecture consisted of an embedding layer, including the vocab size, embedding size and input length corresponding to the max word length.The embedding layer is necessary to transform high-dimensional categorical data into a lower dimensional format allowing the model to learn more easily. Next was a Convolutional 1 D layer to smoothen the input with 32 filters and an activation of relu. A max pooling layer was included, followed by a Bidirectional LSTM layer with 32 neurons. A dropout of 0.4 was chosen for accounting for errors in the model and a final dense output layer of 3, with a softmax activation. The compiling step specified the loss function of ‘categorical \_crossenthropy’, and the optimizer which included the vocab size related to the corpus used to train the model and the learn rate. The model was run over 20 epochs with a batch size of 32. The LSTM achieved an accuracy of 0.9 or 90%. The confusion matrix and model accuracy/loss curves can be seen in Figure 4.10.

Taken together these results indicate that the Extra trees Classifier model had the highest accuracy for the TextBlob sentiment dataset, followed by the XGBoost Classifier and LSTM.



A

B

C

D

E

F

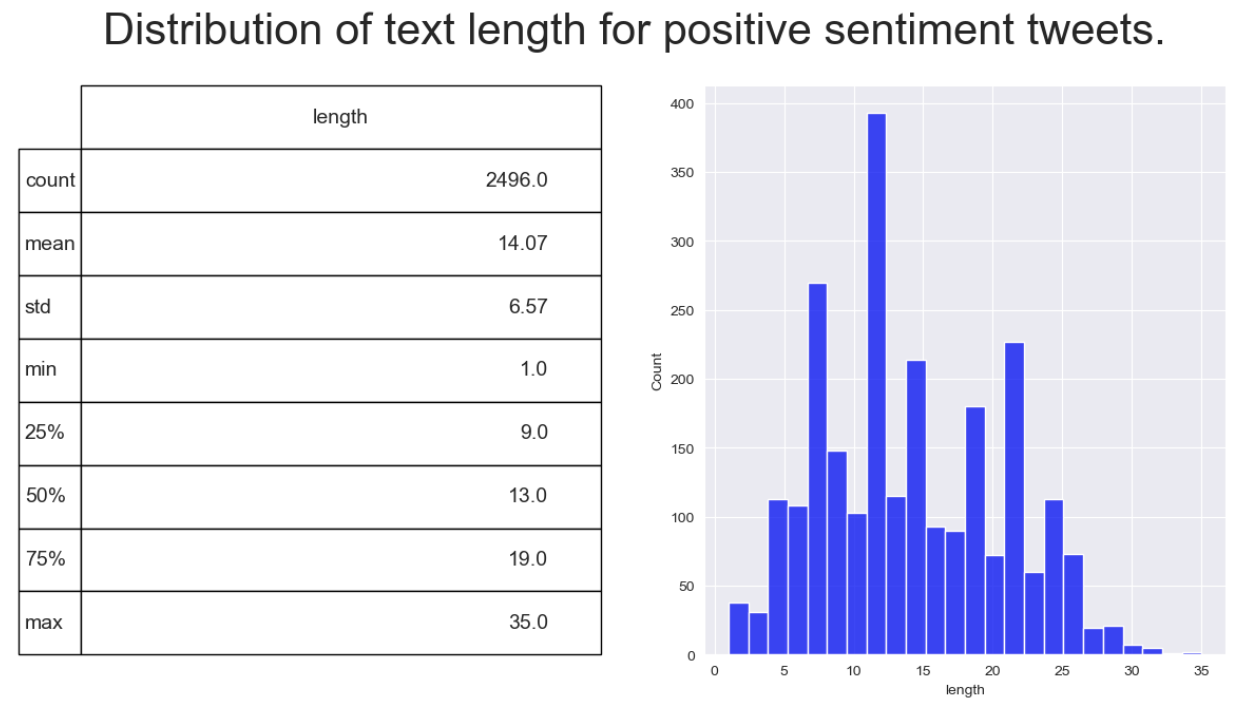
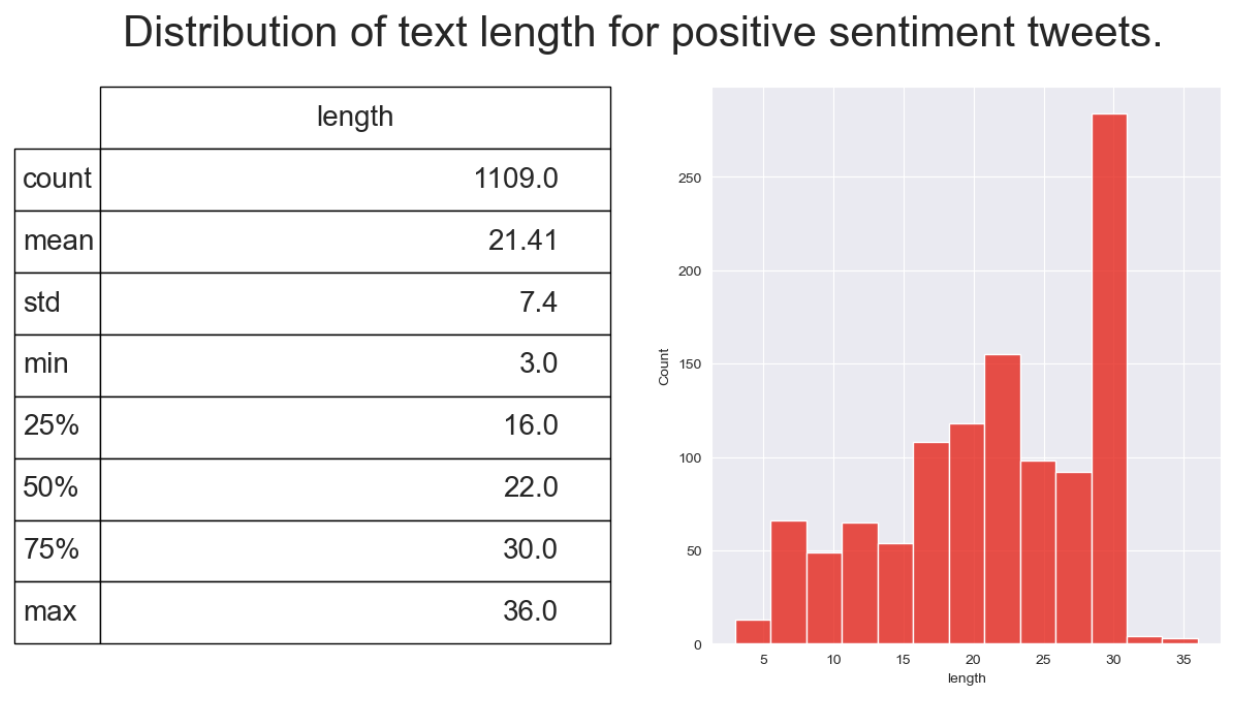
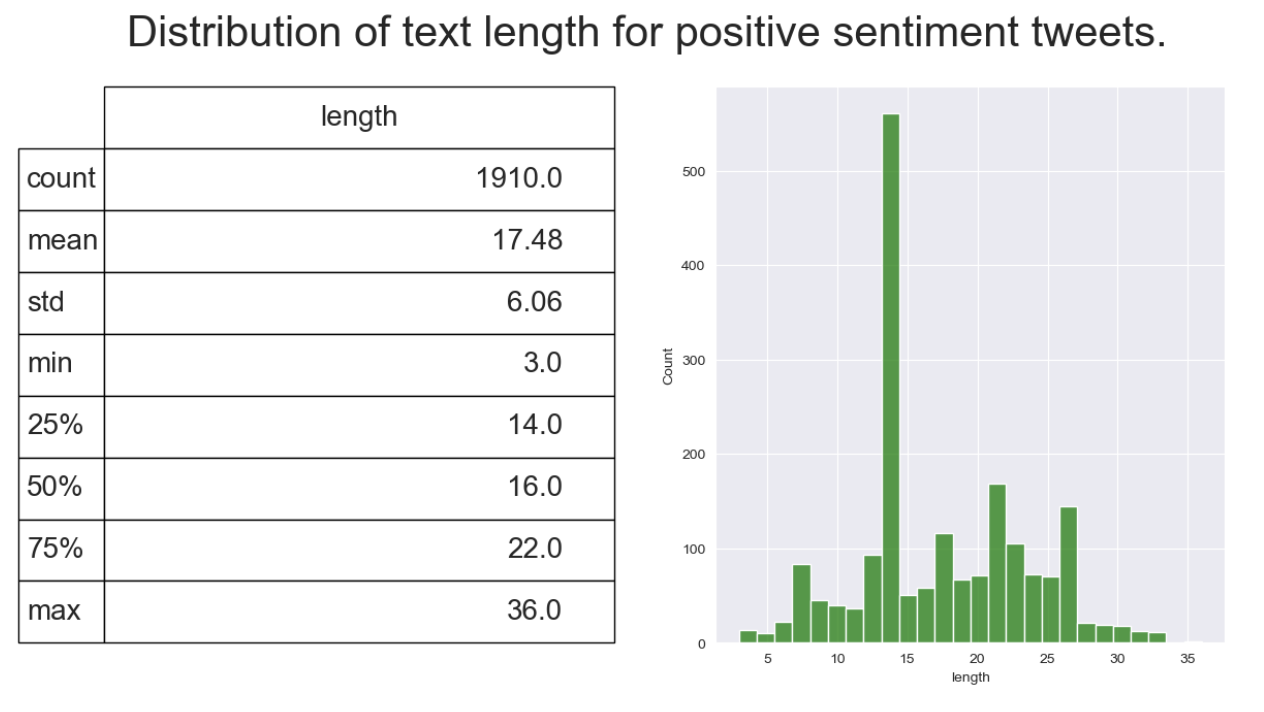
G

Figure 4.7: Confusion matrices showing correctly predicted positive, negative and neutral tweets using different machine learning classification models using TextBlob sentiment.

**A graph showing a graph of a graph

Description automatically generated with medium confidence**

Figure 4.8: Summary plot of model accuracies for classifier models using TextBlob.



A

B

C

Figure 4.9: Distributions of positive, negative and neutral tweet word lengths and descriptive statistics for each sentiment for TextBlob sentiment.

B

A

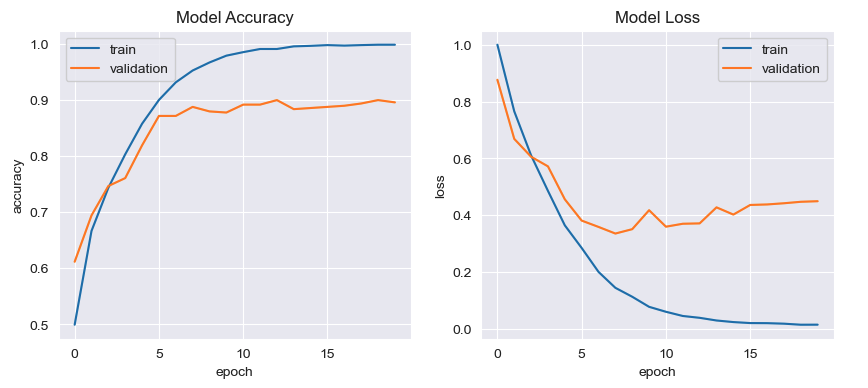
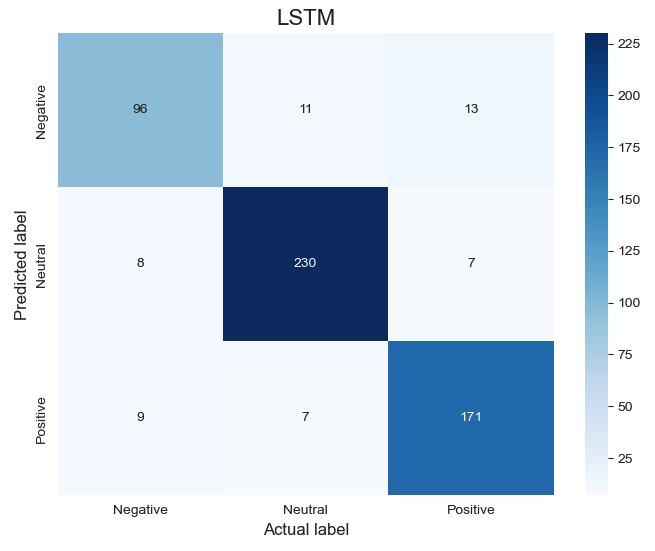
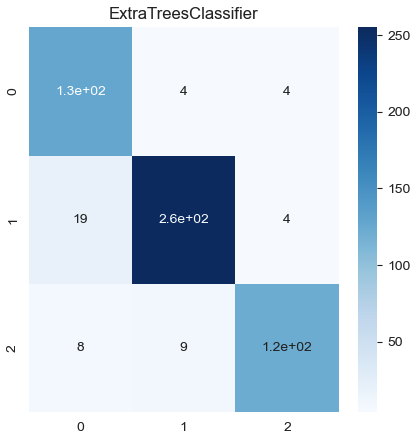
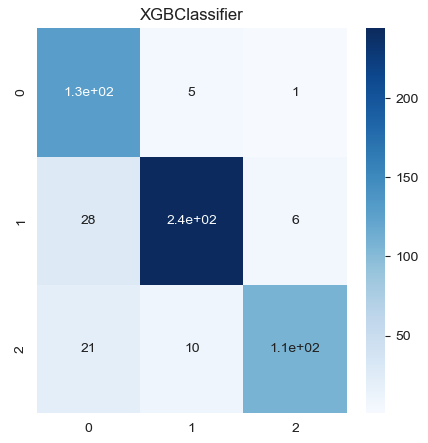
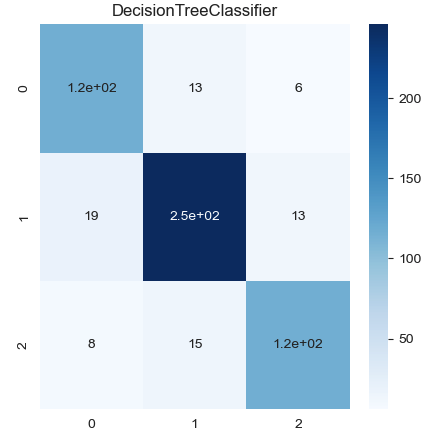
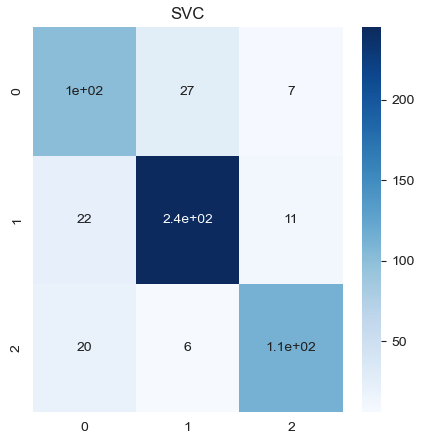
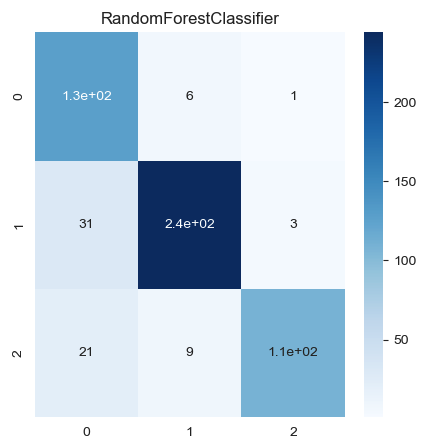


Figure 4.10: Confusion Matrix for correctly predicted sentiment using LSTM and Model accuracy/loss curves for training and validation for TextBlob sentiment.

**4.1.5 Implementation of machine learning models for tweet sentiment classification using VADER sentiment**

In the same manner as Section 4.1.4, the same set of machine learning models were applied to the VADER sentiment dataset to compare model accuracies with the TextBlob sentiment dataset. Models were implemented as previously described in Section 4.1.5. Figure 4.11 shows the confusion matrices for the classification models Random Forest Classifier, Multinomial Naïve Bayes Classifier, Support Vector Machine Classifier, Decision Tree Classifier, K-Nearest Neighbour Classifier, XG Boost Classifier and Extra Trees Classifier. The Extra Tress Classifier was the model with the highest accuracy of 0.91 or 91%. The comparison of model accuracies can be visualised in Figure 4.12. As with the TextBlob sentiment dataset, word length was assessed, shown Figure 4.13, for implementation of the LSTM model. The LSTM architecture was the same as described in Section 4.1.4. The model achieved an accuracy of 0.84 or 84%, 6% lower than the accuracy achieved with the TextBlob sentiment dataset LSTM model. The confusion matrix for predictions of the LSTM with the VADER tweet dataset can be seen in Figure 4.14 , along with the accuracy and loss curves for training and validation. A summary of model accuracies for the TextBlob and VADER sentiment datasets can be seen in Table 4.1. Overall, model performances were similar for each classification model except the LSTM. There are also notable differences in accuracy in the Random Forest Classifier and XG Boost Classifier which will be discussed in Section 5.



A

B

C

D

E

F

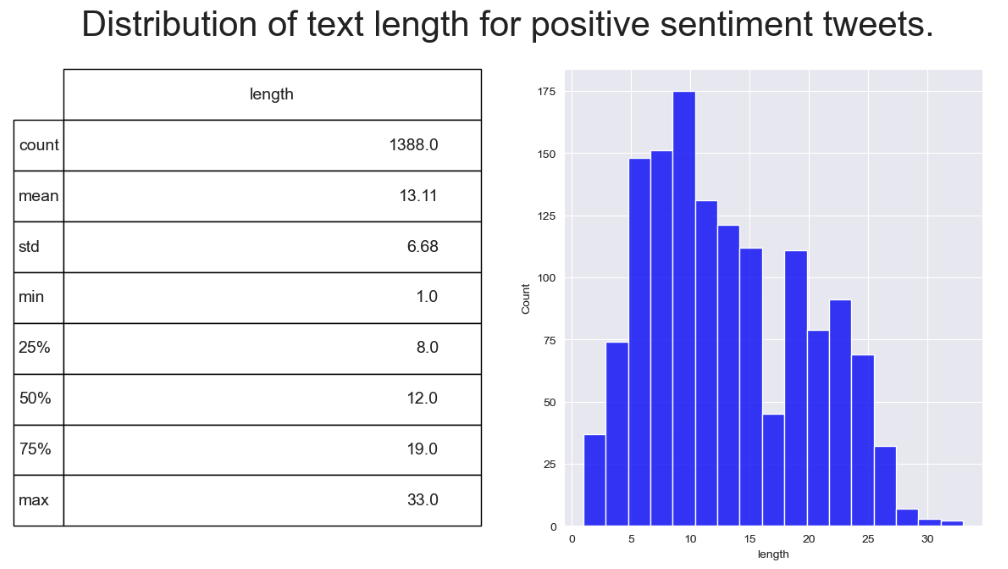
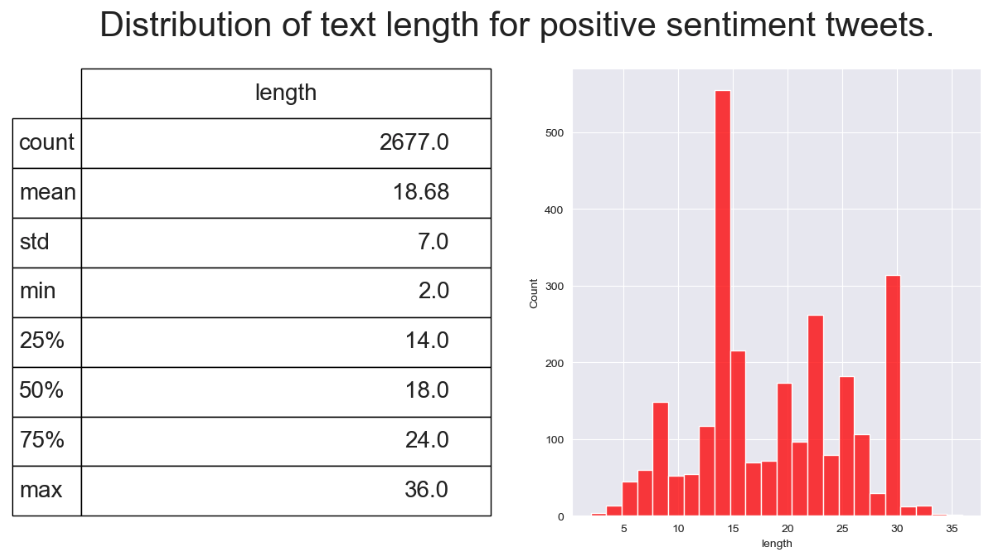
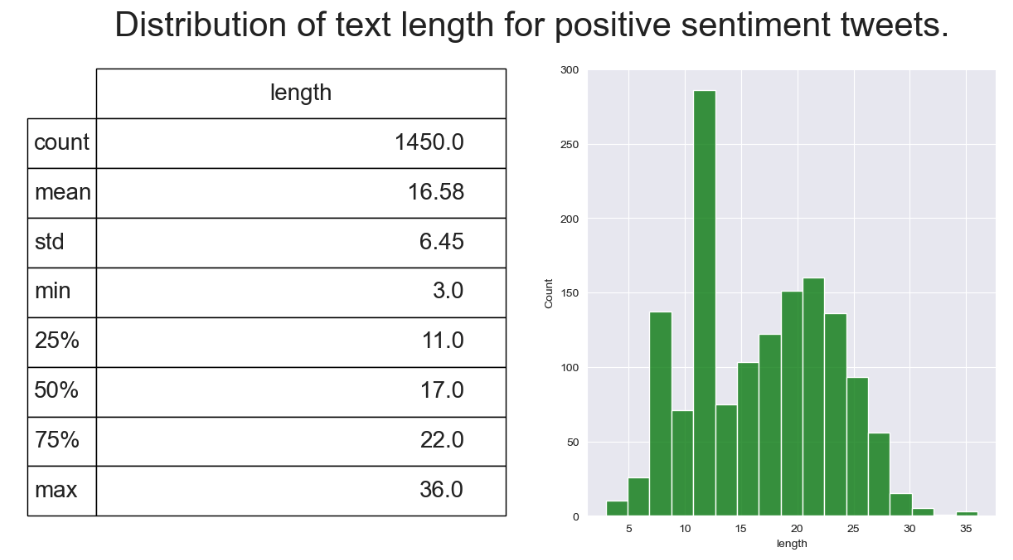
G

Figure 4.11: Confusion matrices showing correctly predicted positive, negative and neutral tweets using different machine learning classification models using VADER sentiment.

**A graph showing a graph of a graph

Description automatically generated with medium confidence**

Figure 4.12: Summary plot of model accuracies for classifier models using VADER.



A

B

C

Figure 4.13: Distributions of positive, negative and neutral tweet word lengths and descriptive statistics for each sentiment using VADER sentiment.

A

B

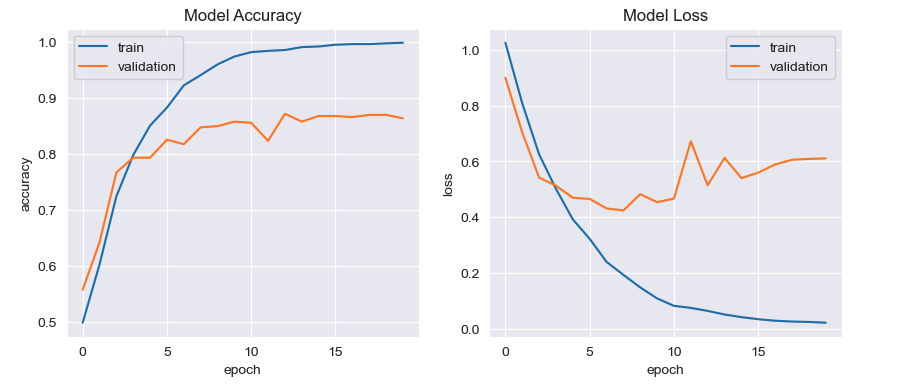
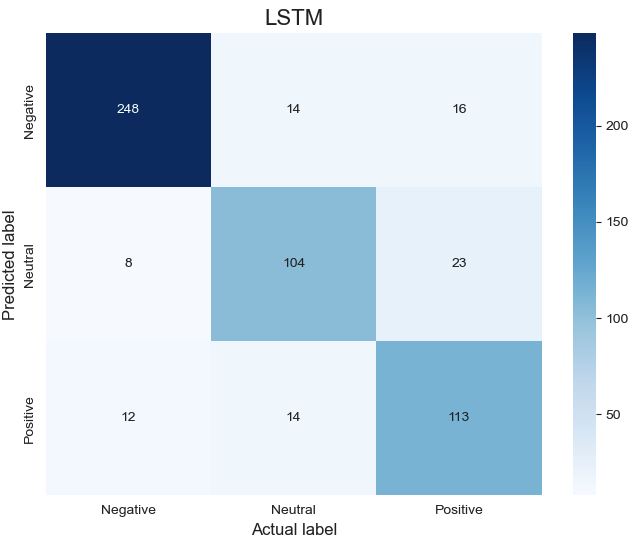


Figure 4.14: Confusion Matrix for correctly predicted sentiment using LSTM and Model accuracy/loss curves for training and validation for VADER sentiment.

****

Table 4.1: Summary table of accuracy, precision, recall and F1-score for tweet sentiment classification models using TextBlob and VADER.

**4.2 Comparative analysis of primary research with covid 19 tweet sentiment**

**A graph showing different colored squares

Description automatically generated**

Figure 4.15: Distribution of Very positive, positive, neutral and negative tweets from Vaccine awareness survey.

**A green circle with a red triangle and a blue triangle

Description automatically generatedA graph of negative emotions

Description automatically generated with medium confidence**

Figure 4.16: Percentages of Vaccine survey sentiment with combined very positive and positive sentiments shown alongside neutral and negative sentiments.

**A graph of a number of bars

Description automatically generated**

Figure 4.17: Comparison of vaccine hesitancy for peoples working in Health/Science areas and non-health/science areas.

**A pie chart with numbers and a red blue and green circle

Description automatically generated**

**A green circle with a percentage

Description automatically generated**

Figure 4.18: Comparison of vaccine sentiment for peoples working in Health/Science areas and non-health/science areas.

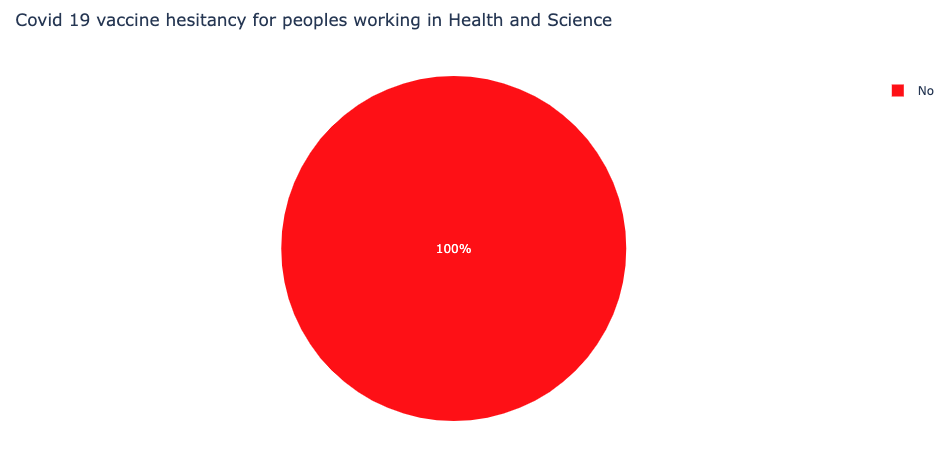
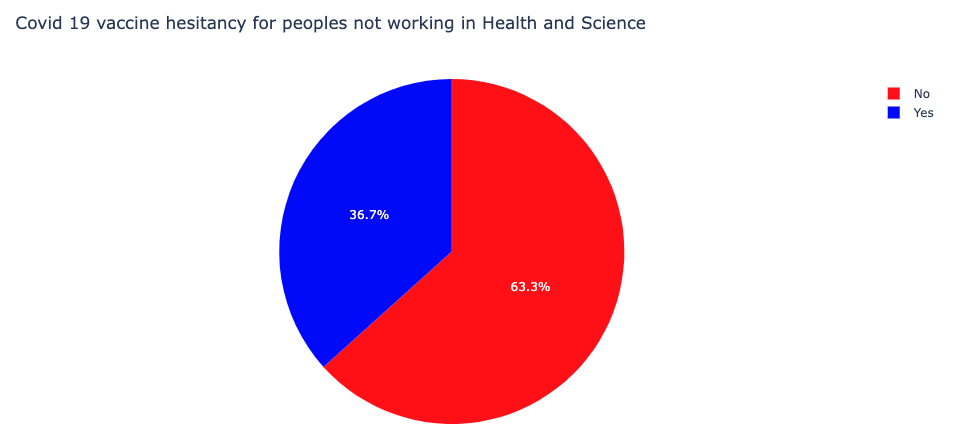


Figure 4.19: Comparison of vaccine hesitancy for peoples working in Health/Science areas and non-health/science areas.

**4.3 Application of forecasting approaches to covid 19 vaccination levels in Ireland**

**5. Discussion (3000 words)**

**6. Conclusions and future work(1000 words)**

**Bibliography**

*A Comprehensive Overview of Sentiment Analysis* (no date). Available at: https://www.analyticsvidhya.com/blog/2022/04/a-comprehensive-overview-of-sentiment-analysis/ (Accessed: 17 May 2023).

*A Sentiment Analysis Approach to Predicting Stock Returns | by Tom Yuz | Medium* (no date). Available at: https://medium.com/@tomyuz/a-sentiment-analysis-approach-to-predicting-stock-returns-d5ca8b75a42 (Accessed: 12 August 2023).

Abonazel, M. R. and Darwish, N. M. (2022) ‘Forecasting confirmed and recovered COVID-19 cases and deaths in Egypt after the genetic mutation of the virus: ARIMA Box-Jenkins approach’, *Commun. Math. Biol. Neurosci.*, 2022(0), p. Article ID 17. doi: 10.28919/CMBN/6888.

Ahmad, W. *et al.* (2022) ‘Enhanced sentiment analysis regarding COVID-19 news from global channels’, *Journal of Computational Social Science*, p. 1. doi: 10.1007/S42001-022-00189-1.

Al-Garadi, M. A., Yang, Y. C. and Sarker, A. (2022) ‘The Role of Natural Language Processing during the COVID-19 Pandemic: Health Applications, Opportunities, and Challenges’, *Healthcare 2022, Vol. 10, Page 2270*, 10(11), p. 2270. doi: 10.3390/HEALTHCARE10112270.

*Beginner’s Guide to XGBoost for Classification Problems | Towards Data Science* (no date). Available at: https://towardsdatascience.com/beginners-guide-to-xgboost-for-classification-problems-50f75aac5390 (Accessed: 27 August 2023).

Binti, A. A., Nasir, M. and Palanichamy, N. (2022) ‘Sentiment Analysis of Covid-19 Tweets by Supervised Machine Learning Models’, *Online) Journal of System and Management Sciences*, 12(6), pp. 50–69. doi: 10.33168/JSMS.2022.0604.

Cascini, F. *et al.* (2021) ‘Developing a Data-Driven Approach in Order to Improve the Safety and Quality of Patient Care’, *Frontiers in Public Health*, 9, p. 667819. doi: 10.3389/FPUBH.2021.667819.

Cepeda, K. and Jaiswal, R. (2022) ‘Sentiment Analysis on Covid-19 Vaccinations in Ireland using Support Vector Machine’, *2022 33rd Irish Signals and Systems Conference, ISSC 2022*. doi: 10.1109/ISSC55427.2022.9826215.

Chandra, R., Jain, A. and Chauhan, D. S. (2022) ‘Deep learning via LSTM models for COVID-19 infection forecasting in India’, *PLOS ONE*, 17(1), p. e0262708. doi: 10.1371/JOURNAL.PONE.0262708.

Chauhan, S. *et al.* (2021) ‘An Artificial-Intelligence-Discovered Functional Ingredient, NRT\_N0G5IJ, Derived from Pisum sativum, Decreases HbA1c in a Prediabetic Population’, *Nutrients*, 13(5). doi: 10.3390/NU13051635.

Chinnasamy, P. *et al.* (2022) ‘COVID-19 vaccine sentiment analysis using public opinions on Twitter’, *Materials Today. Proceedings*, 64, p. 448. doi: 10.1016/J.MATPR.2022.04.809.

*COVID-19 Forecasting and Mathematical Modeling | CDC* (no date). Available at: https://www.cdc.gov/coronavirus/2019-ncov/science/forecasting/forecasting-math-modeling.html (Accessed: 9 August 2023).

Crawshaw, A. F. *et al.* (2022) ‘Defining the determinants of vaccine uptake and undervaccination in migrant populations in Europe to improve routine and COVID-19 vaccine uptake: a systematic review’, *The Lancet Infectious Diseases*, 22(9), pp. e254–e266. doi: 10.1016/S1473-3099(22)00066-4/ATTACHMENT/6EF3885D-1921-4ED3-B114-7351A20CCC97/MMC1.PDF.

Fang, X. and Zhan, J. (2015) ‘Sentiment analysis using product review data’, *Journal of Big Data*, 2(1), pp. 1–14. doi: 10.1186/S40537-015-0015-2/FIGURES/9.

Gecili, E., Ziady, A. and Szczesniak, R. D. (2021) ‘Forecasting COVID-19 confirmed cases, deaths and recoveries: Revisiting established time series modeling through novel applications for the USA and Italy’, *PloS one*, 16(1). doi: 10.1371/JOURNAL.PONE.0244173.

Gonçalves, P. *et al.* (2013) ‘Comparing and combining sentiment analysis methods’, *COSN 2013 - Proceedings of the 2013 Conference on Online Social Networks*, pp. 27–37. doi: 10.1145/2512938.2512951.

Hewage, P. *et al.* (2019) ‘Long-Short Term Memory for an Effective Short-Term Weather Forecasting Model Using Surface Weather Data’, in *IFIP Advances in Information and Communication Technology*, pp. 382–390. doi: 10.1007/978-3-030-19823-7\_32.

Huang, R. and Wang, X. (2023) ‘Impact of COVID-19 on mental health in China: analysis based on sentiment knowledge enhanced pre-training and XGBoost algorithm’, *Frontiers in Public Health*, 11, p. 1170838. doi: 10.3389/FPUBH.2023.1170838/BIBTEX.

Iksan, N. *et al.* (2021) ‘Sentiment Analysis of Public Reaction to COVID19 in Twitter Media using Naïve Bayes Classifier’, *InHeNce 2021 - 2021 IEEE International Conference on Health, Instrumentation and Measurement, and Natural Sciences*. doi: 10.1109/INHENCE52833.2021.9537243.

Ingram, C. *et al.* (2023) ‘Exploring key informants’ perceptions of Covid-19 vaccine hesitancy in a disadvantaged urban community in Ireland: Emergence of a “4Cs” model’, *Vaccine*, 41(2), p. 519. doi: 10.1016/J.VACCINE.2022.11.072.

Iwendi, C. *et al.* (2020) ‘COVID-19 patient health prediction using boosted random forest algorithm’, *Frontiers in Public Health*, 8, p. 562169. doi: 10.3389/FPUBH.2020.00357/BIBTEX.

Khurana, D. *et al.* (2023) ‘Natural language processing: state of the art, current trends and challenges’, *Multimedia Tools and Applications*, 82(3), pp. 3713–3744. doi: 10.1007/S11042-022-13428-4/FIGURES/3.

Kolchyna, O. *et al.* (2015) ‘Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination’. Available at: https://arxiv.org/abs/1507.00955v3 (Accessed: 12 August 2023).

Koroteev, M. V. (2021) ‘BERT: A Review of Applications in Natural Language Processing and Understanding’. Available at: https://arxiv.org/abs/2103.11943v1 (Accessed: 10 August 2023).

Lee, Y. C. and Wu, W. L. (2023) ‘Key Drivers of COVID-19 Vaccine Hesitancy: A Perspective of Collectivism’, *Healthcare*, 11(2). doi: 10.3390/HEALTHCARE11020176.

*Machine Learning Basics with the K-Nearest Neighbors Algorithm | by Onel Harrison | Towards Data Science* (no date). Available at: https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761 (Accessed: 27 August 2023).

Mansoor, M. *et al.* (2020) ‘Global Sentiment Analysis Of COVID-19 Tweets Over Time’. Available at: https://arxiv.org/abs/2010.14234v2 (Accessed: 21 August 2023).

Melton, C. A. *et al.* (2022) ‘Fine-tuned Sentiment Analysis of COVID-19 Vaccine–Related Social Media Data: Comparative Study’, *Journal of Medical Internet Research*, 24(10). doi: 10.2196/40408.

*Most Popular Distance Metrics Used in KNN and When to Use Them - KDnuggets* (no date). Available at: https://www.kdnuggets.com/2020/11/most-popular-distance-metrics-knn.html (Accessed: 27 August 2023).

Müller, A. C. and Guido, S. (2016) *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O’Reilly Media, Incorporated. Available at: https://books.google.ie/books?id=qjUVogEACAAJ.

Müller, M., Salathé, M. and Kummervold, P. E. (2023) ‘COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on Twitter’, *Frontiers in Artificial Intelligence*, 6, p. 1023281. doi: 10.3389/FRAI.2023.1023281/BIBTEX.

*Naives Bayes Classifiers for Machine Learning | by Madison Schott | Capital One Tech | Medium* (no date). Available at: https://medium.com/capital-one-tech/naives-bayes-classifiers-for-machine-learning-2e548bfbd4a1 (Accessed: 26 August 2023).

Qorib, M. *et al.* (2023) ‘Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset’, *Expert Systems with Applications*, 212, p. 118715. doi: 10.1016/J.ESWA.2022.118715.

*Random Forest Algorithm for Absolute Beginners in Data Science* (no date). Available at: https://www.analyticsvidhya.com/blog/2021/10/an-introduction-to-random-forest-algorithm-for-beginners/ (Accessed: 26 August 2023).

*Random Forest Algorithm for Machine Learning | by Madison Schott | Capital One Tech | Medium* (no date). Available at: https://medium.com/capital-one-tech/random-forest-algorithm-for-machine-learning-c4b2c8cc9feb (Accessed: 26 August 2023).

*Regular Expressions — An excellent tool for text analysis or NLP | by Niwratti Kasture | Analytics Vidhya | Medium* (no date). Available at: https://medium.com/analytics-vidhya/regular-expressions-an-excellent-tool-for-text-analysis-or-nlp-d1fa7d666cb9 (Accessed: 17 May 2023).

Ressan, M. B. and Hassan, R. F. (2022) ‘Naïve-Bayes family for sentiment analysis during COVID-19 pandemic and classification tweets’, *Indonesian Journal of Electrical Engineering and Computer Science*, 28(1), pp. 375–383. doi: 10.11591/IJEECS.V28.I1.PP375-383.

Rodríguez-Ibánez, M. *et al.* (2023) ‘A review on sentiment analysis from social media platforms’, *Expert Systems with Applications*, 223, p. 119862. doi: 10.1016/J.ESWA.2023.119862.

Rustam, F. *et al.* (2021) ‘A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis’, *PLoS ONE*, 16(2). doi: 10.1371/JOURNAL.PONE.0245909.

Sarkar, D. (2016) ‘Text Analytics with Python’, *Text Analytics with Python*. doi: 10.1007/978-1-4842-2388-8.

Savage, N. (2021) ‘Tapping into the drug discovery potential of AI’, *Biopharma Dealmakers*. doi: 10.1038/D43747-021-00045-7.

Schröer, C., Kruse, F. and Gómez, J. M. (2021) ‘A systematic literature review on applying CRISP-DM process model’, *Procedia Computer Science*, 181, pp. 526–534. doi: 10.1016/J.PROCS.2021.01.199.

Shahid, F., Zameer, A. and Muneeb, M. (2020) ‘Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM’, *Chaos, Solitons, and Fractals*, 140, p. 110212. doi: 10.1016/J.CHAOS.2020.110212.

Somyanonthanakul, R. *et al.* (2022) ‘Forecasting COVID-19 cases using time series modeling and association rule mining’, *BMC Medical Research Methodology*, 22(1), pp. 1–18. doi: 10.1186/S12874-022-01755-X/TABLES/8.

*Support Vector Machine — Introduction to Machine Learning Algorithms | by Rohith Gandhi | Towards Data Science* (no date). Available at: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47 (Accessed: 27 August 2023).

Syeda, H. B. *et al.* (2021) ‘Role of Machine Learning Techniques to Tackle the COVID-19 Crisis: Systematic Review’, *JMIR Medical Informatics*, 9(1). doi: 10.2196/23811.

*Text Classification with BERT. In this tutorial, we will use BERT to… | by Khang Pham | Medium* (no date). Available at: https://medium.com/@khang.pham.exxact/text-classification-with-bert-7afaacc5e49b (Accessed: 12 August 2023).

*TextBlob | Making Natural Language Processing easy with TextBlob* (no date). Available at: https://www.analyticsvidhya.com/blog/2021/10/making-natural-language-processing-easy-with-textblob/ (Accessed: 17 May 2023).

*TextBlob vs. VADER for Sentiment Analysis Using Python | by Amy @GrabNGoInfo | Towards AI* (no date). Available at: https://pub.towardsai.net/textblob-vs-vader-for-sentiment-analysis-using-python-76883d40f9ae (Accessed: 15 August 2023).

Vaithilingam, S. *et al.* (2023) ‘COVID-19 vaccine hesitancy and its drivers: An empirical study of the vaccine hesitant group in Malaysia’, *PLOS ONE*, 18(3), p. e0282520. doi: 10.1371/JOURNAL.PONE.0282520.

Wang, M. and Hu, F. (2021) ‘The Application of NLTK Library for Python Natural Language Processing in Corpus Research’, *Theory and Practice in Language Studies*, 11(9), pp. 1041–1049. doi: 10.17507/TPLS.1109.09.

*XGBoost: Everything You Need to Know* (no date). Available at: https://neptune.ai/blog/xgboost-everything-you-need-to-know (Accessed: 27 August 2023).

Yamak, P. T., Yujian, L. and Gadosey, P. K. (2019) ‘A comparison between ARIMA, LSTM, and GRU for time series forecasting’, *ACM International Conference Proceeding Series*, pp. 49–55. doi: 10.1145/3377713.3377722.