**A close-up of a logo

Description automatically generated**

**The application of sentiment analysis and forecasting approaches on covid 19 data with focus on Ireland**

by

Gavin Davis, PhD

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 much 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 time series forecasting to understand trends in vaccination sentiment and uptake. Tweet data will be used to generate a predictive vaccine sentiment model with several models being implanted 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. 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. The sentiment will be correlated with other factors collected during the primary research phase. Many factors have been investigated for their contribution to vaccine hesitancy and awareness such as education, ethnicity, religion, age and occupation in other countries. It is the aim of this project to investigate whether there is a contribution of employment background to vaccine hesitancy and awareness in Ireland.

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 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 covid19**.

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 Textblob. Tweets collected will be processed into a new dataset. The likely outcome will be to understand the current state of opinion of vaccine sentiment. In line with this the sentiment data will then be used to train a machine learning model 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 3. Within 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
* 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 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 case/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

**1.3 Research objective 3**

**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

**1.4 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 outline**

**2.2 Datasets**

The tweet dataset used for sentiment mining was generating 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.

**2.3 Collection of Tweet data using Twitter API**

**2.4 Lexicon based sentiment analysis of Tweet data**

**2.5 Machine learning and deep learning models**

**2.4 Proposed 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.

**2.5 Primary research methodology**

The primary research will be performed using a carefully designed and simple questionnaire. The preliminary considerations for the questionnaire design will be what information will be gathered that will help achieve the RO3 of the project, which is to 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 questionnaire 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: Introduction to the survey sent out via email to participants