**A close-up of a logo

Description automatically generated**

**The application of sentiment analysis and time series forecasting to study and develop novel tools for covid19 data**

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

**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 understand trends in vaccination uptake and hesitancy through a number of data analytical approaches. Tweet data will be used to generate a predictive vaccine sentiment model and changes in sentiment over the course of the pandemic will be evaluated. Vaccine uptake will be analysed in Ireland using publicly available datasets from data.gov. The main focus of this project is sentiment analysis and time series forecasting 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 whether there is a contribution of employment background to sentiment around vaccines by carrying out primary research. 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 2 key methods of data analytics as applied to covid19 data, with primary research carried out to answer the hypothesis that employment background may contribute to vaccine uptake and hesitancy:

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.

**2- Research objectives and relevance**

**2.1 Research objective 1**

**The aim of this research objective is to investigate vaccine sentiment**.

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. The polarity will be analysed and visualized to insights into the current opinion of vaccines. 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 for comparative analysis. 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 and other parameters related to covid 19.

**2.2 Research objective 2**

**The aim of this research objective is to investigate whether time series forecasting can be applied to covid19 case numbers and/or vaccine levels to generate accurate forecasting models.** To carry out this aspect of the research project data will be used from data.gov on case and vaccination levels and case numbers from Ireland. 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. Time series analysis models such as LSTM, RNN and ARIMA will be used to predict the chosen parameter with each models suitability analysed. The likely outcome of this part of the research will be the generation of an accurate predictive model which could assist the understanding and prediction of case/vaccine levels in Ireland during the pandemic over time. In performing this analysis a comparison of different counties in Ireland will be made in terms of case levels and vaccine uptake. 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 time series analysis and the models which have been applied to covid19 data.

**2.3 Research objective 3**

**The aim of this research objective is to investigate whether there is a contribution of employment background on vaccine hesitancy and awareness, and whether awareness around vaccines has increased in employment subgroups since the pandemic**. 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 used and compared to the sentiment analysis collected as a part of RO1 to understand whether the sentiment from twitter data and that of the people of Ireland is similar or different.

This data will be correlated with employment profile to understand whether there is a relationship between employment and sentiment towards vaccines post pandemic. The two main subgroups in question are people working in science/health and people working in non-science. 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 .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.

**2.4 Relevance of project**

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 techniques 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. The application of time series forecasting to highlight models with accurate predictive power for covid 19 vaccine levels and case levels will showcase which models are best for future use in the prediction of such metrics.

**Literature review**

**Methodology**

**Sampling**

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. A detailed method and justification for the primary research approach will be given in a following section.