National College of Ireland

Using artificial intelligence techniques to analyse social media content on COVID-19 children vaccination programs

MSc Research Project
Master of Science in Data Analytics

Anne Guilcher
X16132068

School of Computing
National College of Ireland

Supervisor: Vladimir Milosavljevic

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Anne Guilcher
Student ID:	X16132068
Programme:	Master of Science in Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Vladimir Milosavljevic
Submission Due Date:	15/08/2022
Project Title:	Using artificial intelligence techniques to analyse social media content on COVID-19 children vaccination programs
Word Count:	7,221
Page Count:	28

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Anne Guilcher
Date:	14th August 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Using artificial intelligence techniques to analyse social media content on COVID-19 children vaccination programs

Anne Guilcher
X16132068

Abstract

The vaccination of children against the COVID-19 virus was first authorised in the United States in May 2021. Thereafter, people worldwide started to express their personal feelings towards COVID-19 vaccination programs with initiatives aimed specifically at children becoming a hotly debated and highly divisive topic. As information posted on social media platforms such as Twitter is publicly accessible and can be extracted to identify opinions, sentiment, trends and patterns, the purpose of this particular paper is to analyse the variety of opinions specifically linked to COVID-19 vaccination programs amongst children. Classical machine learning models and deep learning algorithms were implemented and correlated to establish the most efficient classifier. 1,019,661 tweets have been gathered, analysed and correlated against key events identified during the rollout of COVID-19 vaccines amongst children as reported by various media outlets. From the analysis undertaken, it has been observed that the majority of the tweets related to COVID-19 vaccination programs of children are neutral, whilst the number of tweets in favour of such programs outnumbers those which expressed negative sentiment. In addition, the findings of this present research have highlighted that the frequency of posts linked to the vaccination of children against COVID-19 follows the timeline and trend of news events. This research paper involved the development and comparison of 12 classifiers with the most optimal model being BERT with an accuracy rate of 90.3%. The proposed approach could similarly be used to monitor existing vaccination programs to help governments better design appropriate communication campaigns and assist them in offering the public clear, detailed and targeted information. In addition, such research findings could support decision makers and health professional to identify and address public safety concerns so as to effectively debunk misinformation and conspiracy theories. Finally, this research initiative could potentially act as a catalyst to help influence public attitudes towards COVID-19 children immunisation programs and increase overall public trust in future vaccination rollouts thereafter.

1 Introduction

In March 2020, the World Health Organization (WHO) declared that the COVID-19 virus had become a worldwide pandemic. This pandemic has subsequently become one of the most unprecedented challenges in modern times unleashing disastrous effects on both

economic activity and public health globally. The successful development of COVID-19 vaccines has been instrumental in combatting this deadly virus as the primary goal of public health authorities subsequently shifted to the rapid delivery of effective vaccine programs. Improving vaccine rollouts has been imperative in the response to COVID-19 as a significant cross-section of the population require inoculation in order to achieve herd immunity. As public sentiment towards COVID-19 vaccines significantly impacts the successful outcome of vaccination campaigns, interpreting patterns and trends on social media platforms is vital to policy makers when designing vaccination strategies. While social media is frequently used to share information regarding COVID-19, it also facilitates easy access to incorrect information and misinformed opinions. Several studies like Dunn et al. (2015) and Kornides et al. (2022) have demonstrated that exposure to misinformation, rumours and negative sentiment on Twitter has increased scepticism towards vaccines in general thus resulting in a decline in vaccination uptake.

Research Question and Objectives
Understanding the public's concerns regarding COVID-19 children vaccination programs is an important component in achieving optimal vaccine rollout as successfully targeted information and educational campaigns will result in enhanced public trust towards policy makers and health organisations initiatives. This research paper has focused on the following question and sub-question:

RQ: "How can artificial intelligence techniques be used to analyse media content to give insight into public sentiment attached to COVID-19 children vaccination campaigns?" and SUB-RQ:"How well can classical and deep learning algorithms classify the sentiment categories attached to the vaccination of children against COVID-19?".

The goal of this particular research is to gain a better understanding of worldwide public opinion and sentiment towards COVID-19 children vaccination programs. Based on research findings thus far, this appears to be the first study to use data extracted from Twitter to better understand public opinion regarding COVID-19 vaccination programs amongst children. Given the significant influence of Twitter as a social media platform, this research project involved an analysis of tweets during the timeframe between February 2020 and June 2022. In order to gain a deeper insight into the public perception of COVID-19 child vaccination programs, this research paper did not focus on one specific location but did however limit its analysis to Twitter posts conducted in just one language - English. The objective of this study is to determine public sentiment towards child immunisation, identify the main areas of concern and misinformation as well as investigate multiple classical and deep learning algorithms so as to determine the most efficient model.

2 Related Work

Information, opinions and concerns about COVID-19 have been widely shared on social media platforms since the beginning of the pandemic. Such platforms quickly became the fastest channel of communication and played a crucial role in keeping the public informed during successive lockdowns. However, they also contributed to the dissemination of misinformation and negative vaccination content which may contribute to increased vaccine hesitancy.

News and social media roles during crisis

There was a noticeable increase in the usage of social media throughout successive lockdowns as the public relied heavily on such platforms to share information regarding daily new cases numbers, infection rates, mortality statistics, vaccine rollout progress and general government guidelines. Multiple challenges emerged as increased irresponsible online activity resulted in higher levels of anxiety and a deeper distrust towards decisions makers (González-Padillal and Tortolero-Blanco (2020)).

Clear and coherent communication from public health bodies has been an important component in implementing successful strategies during the COVID-19 pandemic. Integrity, transparency and trust are key characteristics which must be prevalent amongst policy makers while articulating vaccination programs. Social media platforms are pivotal in formulating clear communication and supporting crisis management during emergencies (Tang et al. (2018)). Merchant and Lurie (2020) demonstrated a noticeable link between increased frequency of tweets during the evolution of the pandemic as official figures were released by the Center for Disease Control and Prevention. The findings of this particular study concluded that Twitter could assist in producing real-time estimation and predictions during pandemics. Following on from that research, Kaur et al. (2020) similarly studied data extracted from Twitter to monitor emotions and sentiment trends during the first few months of the COVID-19 pandemic.

Data mining and analysis of social media content are useful techniques when attempting to better understand public emotions. Once a more comprehensive understanding of public sentiment is acquired, policy-makers and health organisations can communicate more efficiently with its citizens.

Data mining with artificial intelligence techniques

Data can be extracted from multiple online sources. Ruiz and Bell (2021) and Piltch-Loeb et al. (2021) used online-based questionnaires to research the public's reluctance towards vaccination programs through the use of multiple linear regression. These research initiatives all encountered similar challenges and limitations linked to the manner in which data was collected such as respondents unequal repartition and inadequate population representation. Additional research undertaken utilised commercial databases (Evanega et al. (2020), Martin et al. (2020)) and web scraping (Krawczyk et al. (2021), Rahul et al. (2021)) which usually allow for the gathering of much larger data sets.

Furthermore, scraping online posts allows for text as well as metadata analysis similar to that carried out in Abd-Alrazaq et al. (2020) and Krawczyk et al. (2021). Such studies analysed tweets metadata such as re-tweets to enhance their models to highlight and investigate the main themes discussed, authors' concerns, emotions, sentiment and public interaction. Once information from media platforms has been extracted, it must then be linguistically understood through text mining approaches. Text mining techniques allow for linguistic features to be extracted which is a necessity prior to the analysis. Various types of syntax features are examined such as the average length of tweets, word counts, number of characters used, words patterns and complexity (Boucher et al. (2021)) as well as lexicon and semantic features as demonstrated by Rahul et al. (2021). Text mining techniques like Natural Language Processing (NLP) in association with supervised or unsupervised models are widely used to exploit text-content data, extract topics and analyse and predict sentiment. Similarly NLP methodology is widely used to analyse and classify sentiment attached to social media posts, be they negative, neutral or posit-

ive. Artificial intelligence techniques use the National Research Council (NRC) Emotion Lexicon to evaluate emotions. TextBlob and Valence Aware Dictionary for Sentiment Reasoning (VADER) are both examples of lexicon-based techniques utilised in Hussain et al. (2021) to evaluate sentiment attached to COVID-19 vaccines on two separate social media platforms. Similarly Shapiro et al. (2017) and Luo et al. (2019) used NLP to analyse sentiment regarding the HPV vaccination. While Shapiro et al. (2017) used the Health Belief Model (HBM) for tweets, amount of followers and classification supervised machine learning techniques to investigate social connections, the results produced were somewhat restricted in that the study did not utilise other social interactions to evaluate the spread and influence of negative opinions attached to vaccination. In addition, the corpus size and a lack of fine-tuning in the threshold defined to identify extremely positive or negative tweets limited this study's findings. Links between content attached to vaccines and Twitter users' engagement were also investigated in Shamrat et al. (2021). Whilst over-dispersion of retweet frequency in the data set was compensated through the use of negative binomial regression models, the sample size of this particular study was relatively small (two hashtags were used while creating the corpus) and also limited to the English language. Although the research papers listed above had a variety of techniques implemented, the findings may not be considered as representative of the entire public at large as they all predominantly used data extracted from one single online media source - Twitter.

Topic modelling and assessment techniques

Data can be organised in different themes by using topic modelling which in turn can be analysed further by using metrics such as perplexity and coherence (Lyu et al. (2021)). In addition, the performance of a model can be enhanced by amending the amount of latent topics to be retrieved. Studies such as (Lyu et al. (2021), Abd-Alrazaq et al. (2020) and Hussain et al. (2021)) have analysed the most recurrent themes and trends linked to COVID-19 on social media platforms with these research papers using unsupervised machine learning methods such as Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA) algorithms as well as rule-based classifications. However, these studies were somewhat limited in their findings due to the size of data sets utilised (too big or too restricted), the use of very few hashtags (Xue et al. (2021)) as well as their sole focus on the English language.

Information extracted from multiple social media platforms such as Twitter, Instagram or Reddit was analysed and compared by Cinelli et al. (2021) where NLP techniques and the Partitioning Around Medoids (PAM) algorithm were implemented to evaluate themes. Nevertheless, this particular study encountered some challenges linked to the filtering applied to the extracted data, which greatly reduced the size of the sample. Enhancing filtering techniques would have greatly improved the model developed.

The level of untrustworthy websites mentioned in tweets and their possible influence were also assessed in other research papers (Shah et al. (2019)) where multiple classifiers (random forest, support vector machine and recurrent neural network) were developed and trained on the scraped data. However, some limitations to this study emerged as other types of models such as logistic regression and naïve bayes were not investigated. Furthermore, the size of the sample used was small which resulted in training data of poor quality.

Deep learning techniques
Many studies have demonstrated that results obtained through deep learning (DL) based classifiers are more accurate. Deep learning can support the identification of unfounded concerns and the source of misinformation when coupled with social network analysis. Such techniques are crucial in the development of successful vaccination campaigns. Deep learning models can be configured with numerous parameters and different configurations were used in the model design of the following selected studies. Cinelli et al. (2021) utilised the skip-gram neural networks model on sizeable data sets to implement word embedding as this method outperformed Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) on very large corpus.

Convolutional Neural Networks (CNN) and Artificial Neural Networks (ANN) were implemented in Reddy and Reddy (2019) and Cotfas et al. (2021) to analyse sentiment in tweets and opinions.

Long-Short Term Memory (LSTM) was utilised in Paul and Gokhale (2021) and Bi-directional LSTM (Bi-LSTM) in the Cotfas et al. (2021) and To et al. (2021) studies to analyse and classify opinions and sentiment on vaccines. Bidirectional Encoder Representations from Transformers (BERT) model is a transformer-based machine learning method which was specifically developed for NLP. This method was implemented in numerous studies on sentiment attached to COVID-19 and vaccines in general. All of these selected studies resulted in well-performing models. The following research papers highlighted a higher accuracy with the BERT method than with other models: (Hussain et al. (2021)), Garcia and Berton (2021), Müller et al. (2020) and Cotfas et al. (2021). Similarly, other papers such as To et al. (2021) developed their own implementation of the BERT model and produced results with marginal improvement of 10-30% compared to the base model.

In this context, the objective of this research paper is to analyze Twitter users perception and sentiment attached to COVID-19 children vaccination programs. Whilst previous research focused mainly on data related to COVID-19 vaccines in general, this present study has investigated a social media platform (Twitter) to analyse emotions and themes linked to COVID-19 children vaccination campaigns through the use of supervised machine learning and deep learning techniques. In addition, the research conducted has analysed data in a global context as opposed to one specific geographic location and has focused on information communicated through the English language.

3 Methodology

3.1 COVID-19 Children Vaccination Programs Methodology

A new research framework called "Methodology for COVID-19 Children Vaccination Programs" was developed specifically for this project to analyse social media content of COVID-19 Children Vaccination initiatives using natural language processing (NLP) techniques. It is a tailored version of the Cross Industry Standard Process for Data Mining (CRISP-DM) and is detailed in Figure 1. This iterative approach is based on five of the six CRISP-DM stages and allows for regular opportunities to assess progress, minimise risk and ensure the project's objectives are always focused on.

Research statement understanding
This step involved the clear definition of objectives and challenges linked to extensive data mining, manipulation and analysis in order to gain a good understanding of the

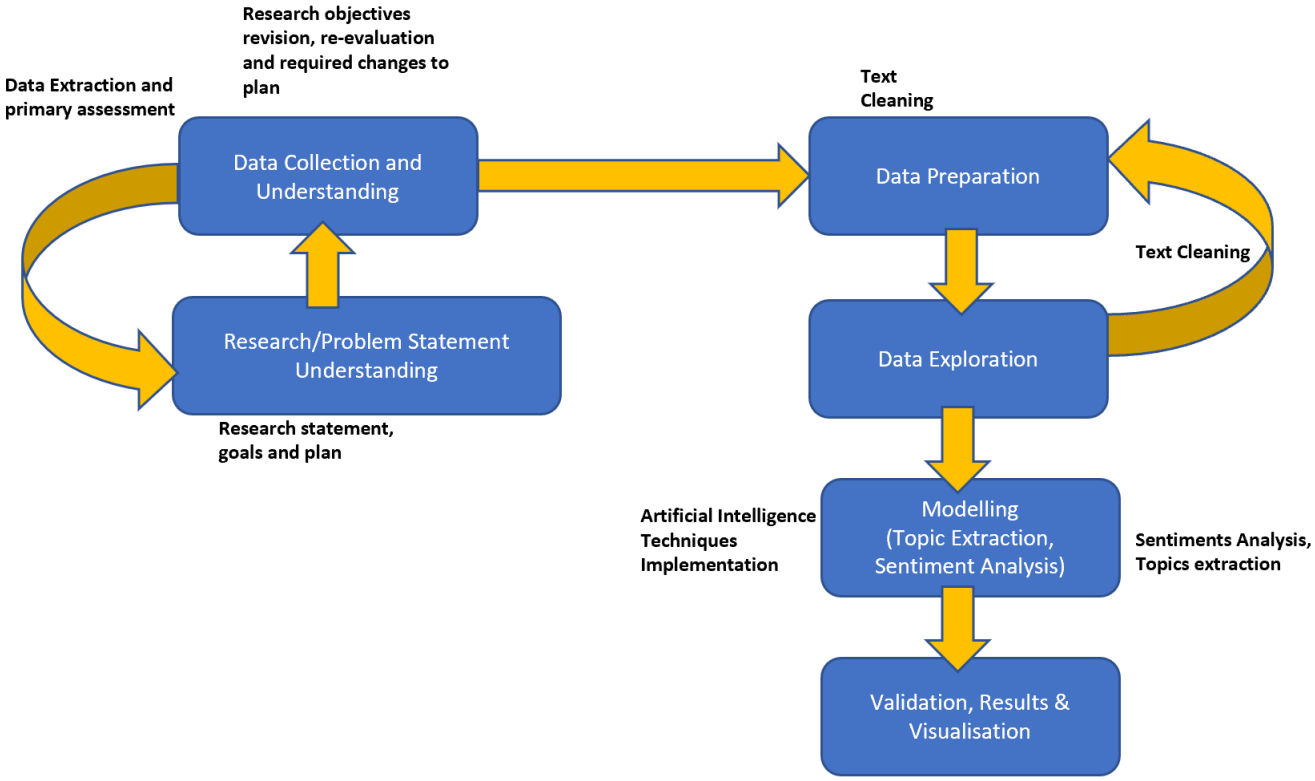

Figure 1: COVID-19 children vaccination programs methodology

research/business statement. A clear understanding of the project scope is crucial to the success of any research project. This particular research initiative involved the extraction and analysis of a large corpus from Twitter specific to the issue of the vaccination of children against COVID-19. This required an extensive use of NLP techniques to highlight topics, analyse sentiment as well as the creation and comparison of multiple models to predict sentiment attached to tweets. In addition, the research incorporated analysis of the tweets content including the impact of factors such as total number of cases, death or vaccination.

Data collection and understanding

Data specific to COVID-19 children vaccination programs was collected from a single social media source (Twitter).

A second dataset was retrieved from Our World in Data website [1] (attached to the University of Oxford) to retrieve daily statistics on COVID-19 pandemic in 207 countries. The type of information retrieved included daily total cases, new cases, total deaths, new deaths, reproduction rate, total vaccinations as well as total boosters rates. In addition, demographic data such as population density, median age, aged over 65, over 70, GDP per capita, extreme poverty, cardiovascular death rate, diabete prevalence, female and male smokers, handwashing facilities, number of hospital beds, life expectancy as well as the human development index were also included. The data was subsequently filtered (only statistics related to location 'World' were kept) and merged with the tweets data set based on the date columns in both data sets.

[1] Our World in Data: `https://ourworldindata.org/coronavirus`

Data preparation
This particular phase of the research is a crucial step in advance of modelling as its objective is to transform the raw text-content extracted from Twitter into exploitable data to carry out analysis and ultimately extract significant insights. To do so, data cleaning is implemented through the use of relevant techniques in Python such as the removal of punctuation, stop words, special characters, URLs or the conversion to lowercase. This data transformation step is then followed by a data exploration phase.

Data exploration
This phase involves the initial analysis of the retrieved data through the creation of multiple graphs and visuals. Continuous manipulation and cleaning of the data, based on the exploratory analysis initial findings, are required to create a data set which can then be used in the modelling phase. As recommended by V. Raghupathi, Ren, and W. Raghupathi 2020, word count frequency and word co-occurrence analysis were implemented. As the scraped data set is large, a random subset of the transformed data was first extracted and then resampled prior to the development of models. The random extract consists of 153,096 tweets and is chronologically organised as follows: 9,720 (for 2020), 109,383 (for 2021) and 33,993 (for 2022).

Modelling
Multiple models, selected based on findings from existing literature and research, were developed throughout this study. NLP techniques in particular were extensively researched and implemented. Content polarity was incorporated through the use of the NCR lexicon-based methodology. Text-content from tweets was analysed and sentiment extracted with the Valence Aware Dictionary for Sentiment Reasoning (VADER) technique. Latent Dirichlet Allocation (LDA) was used to model topics as recommended by Rahul et al. (2021). In addition, the following classical models were implemented to analyse and predict sentiment and their results assessed and compared: logistic regression, random forest, extreme gradient boost, K-Nearest Neighbour, decision tree and naive bayes. Furthermore, deep learning techniques such as simple RNN, single LSTM layer, bidirectional LSTM, 1D convolutional, Bi-directional Encoding Representation for a Transformer (BERT) as implemented by Alam et al. (2021) and DistilBERT were developed. All of these various models were fine-tuned to enhance their performance.

Validation and results
The models were not deployed in a separate environment. Their evaluation involved the analysis of multiple metrics such as precision, accuracy, F1 scores and these results were subsequently compared to assess and rank the models.

3.2 Text Analysis Pipeline for COVID-19 Children Vaccines

The method implemented to analyse text-content throughout this project is composed of three stages: collection of data, data pre-processing and analysis of text as shown in Figure 2.

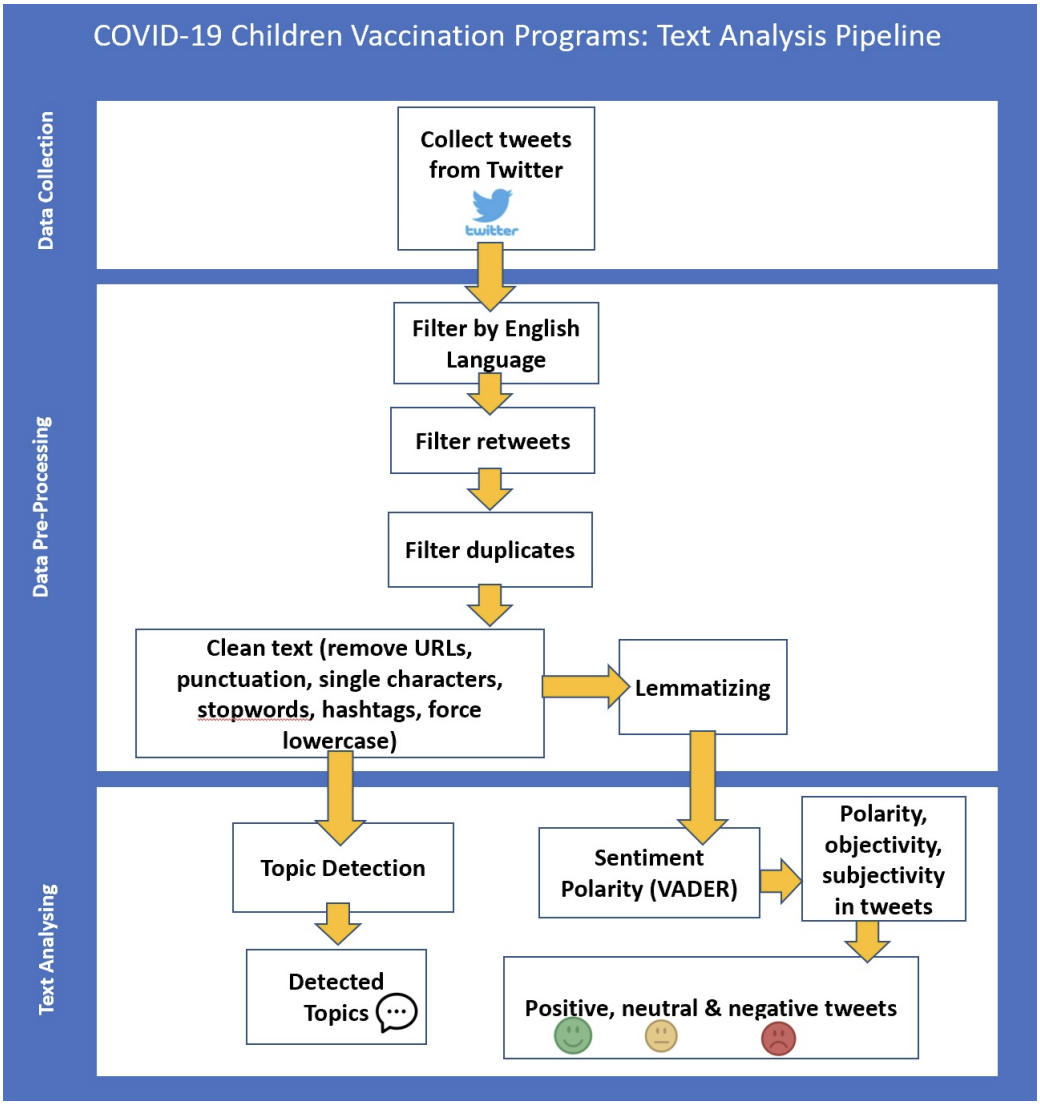

Figure 2: Text analysis pipeline for COVID-19 children vaccination programs

3.2.1 Data extracted from Twitter

Python was extensively used to collect, format, clean and transform data as necessary to produce a corpus for the text-based emotion and sentiment analysis required to meet this research's objectives.

Data collection

In terms of information gathering, data was scraped from a unique social media source (Twitter) using Python, the Twitter API and twarc library. The establishment of an academic research Twitter account allowed for the extraction of a large collection of tweets between February 2020 and June 2022. Dates were selected to allow for the extraction of a wide cross section of tweets as well as to facilitate the merge of the Twitter data set with the data set extracted from 'Our World in Data' website [2]. A query was developed to extract tweets written in the English language using the following keywords: corona,

[2]Our World in Data: `https://ourworldindata.org/coronavirus`

coronavirus, covid-19, covid, kids, child, children, kid, vaccination, vaccine, vax. Monthly data was subsequently extracted and output into json files through the use of Python, the Twitter API and twarc library in a Jupyter notebook.

Data formatting
Using the Pycharm IDE, further Python code was implemented to transform each json file into a csv file. This was followed by a high-level analysis of the data. As the scraped files initially had 73 columns, a preliminary cleaning phase took place where fields relevant to the project's objectives were retained while others were deleted from each data set [3]. Afterwards, all the csv files were concatenated into a single one [4]. The overall tweets data set holds 1,019,661 tweets across 27 columns. As a large proportion of the tweets authors' location was undefined or unexploitable [5], the study analysed tweets at a worldwide/global level and was not specific to a particular region.

Data pre-processing
Quality of text is key to ensuring data insights extracted during the analysis are meaningful. This research involved an extensive cleaning and manipulation of tweets text-content with case folding implemented to convert text to lowercase. URLs, punctuation, emojis, emails, new lines, white spaces, single characters, special characters, hashtags and single quotes were all removed whilst stop words (which are irrelevant or potentially irrelevant) were also deleted. Slang and words contractions were processed as well. Text was tokenized for further processing thereafter. In addition, text was lemmatized to standardise words.

3.2.2 Data extracted from 'Our World in Data' website

A csv file holding numerous daily metrics regarding the COVID-19 pandemic (starting from February 2020 through to June 2022) was manually downloaded from the 'Our World in Data' website. The initial data set contained 69 columns and 197,003 rows. This data set was subsequently filtered and a new data set created storing only global/worldwide statistics. This final data set ultimately held 45 columns and 887 rows. After carrying out extensive text-content analysis and modelling on the tweets data, the two data sets were merged and various models developed and assessed to investigate the influence some external predictors might have on the sentiment attached to Twitter posts [6].

[3]Refer to Configuration Manual, sections 3.3 & 3.6
[4]Refer to Configuration Manual, section 3.4
[5]Refer to Configuration Manual, section 3.8
[6]Refer to Configuration Manual, section 3.17

4 Design Specification

A 3-tier architecture software design approach was implemented in this research project with presentation, business logic and persistence layers as depicted in Figure 3.

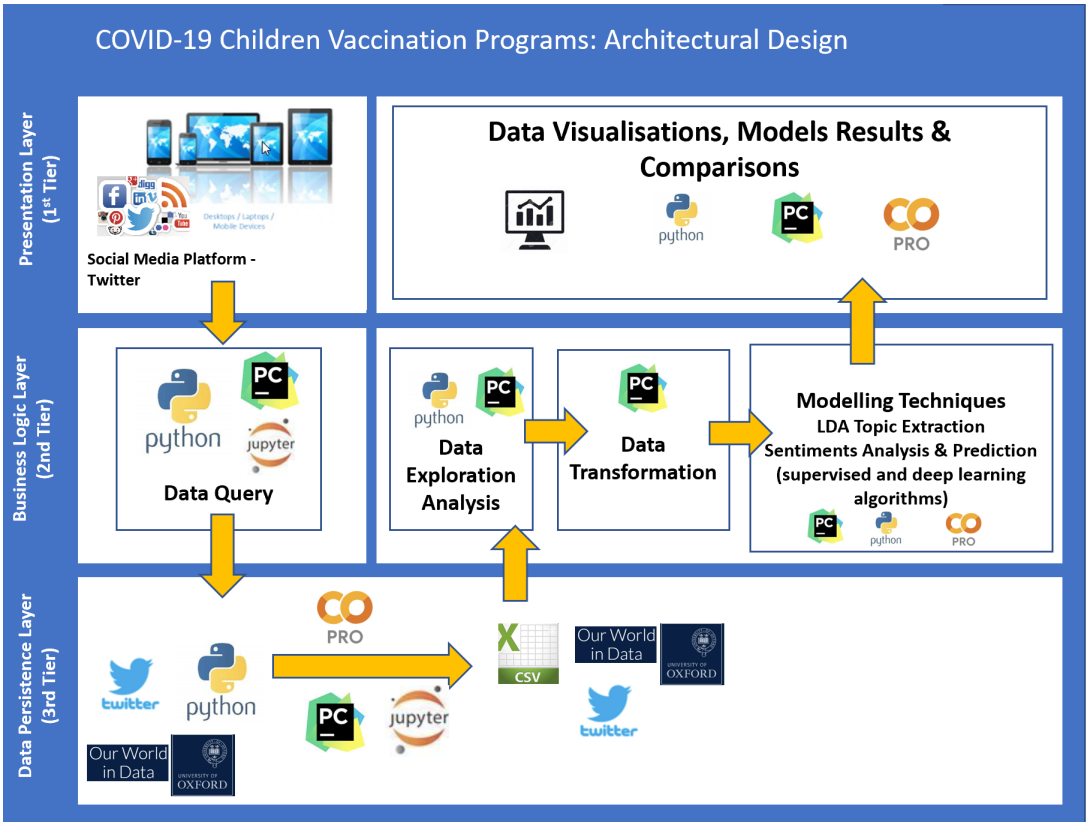

Figure 3: COVID-19 children vaccination programs design specifications

The presentation/client layer represents the user interface where results are presented to stakeholders. Multiple NLP techniques are implemented in the business logic tier to process data, carry out exploratory analysis, highlight topics, analyse sentiment and implement numerous models. Finally, the data itself resides in the third tier: the data persistence layer. Python programming language and relevant libraries were extensively used in this project to extract, clean, process, analyse data, implement, assess and compare models as well as present results. Furthermore, the use of numerous tools (Tableau for initial data visualisations, SPSS for initial data analysis, Jupyter Notebook, PyCharm and Google Colab Pro for extensive analysis and modelling) facilitated the text-content analysis detailed in the next section.

5 Implementation

A random extract of the tweets data set was used to carry out an analysis and modelling of sentiment attached to COVID-19 children vaccination programs. In addition, a class imbalance issue was resolved by upsampling the data. This new data set holds 153,096 rows.

5.1 Exploratory Analysis

Python was chosen as the language to perform an exploratory analysis to gain further insights on the data prior to extracting topics, sentiment analysis and apply modelling techniques.

5.1.1 Analysis of number of tweets

The number of tweets posted related to the vaccination of children against COVID-19 was graphed and analysed. Multiple graphs were developed to depict the number of tweets posted daily and monthly [7] with Figure 4 outlining the number of tweets per month between February 2020 and June 2022.

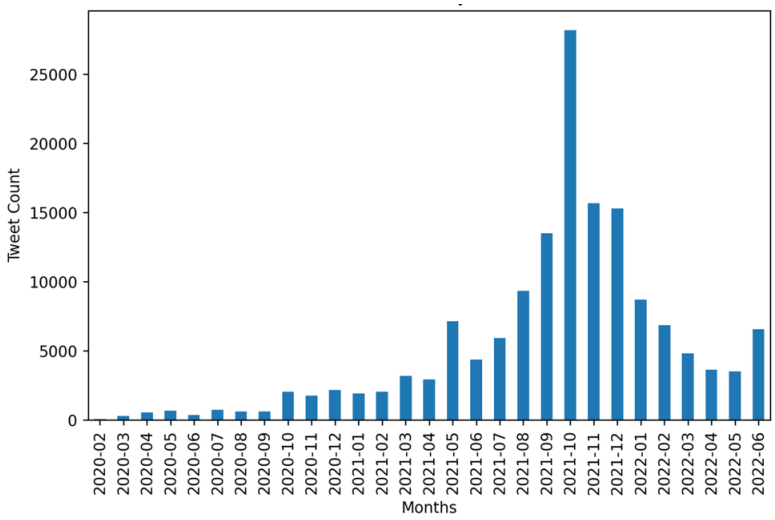

Figure 4: Number of tweets per month

The number of tweets per month is detailed further in Table 1. Whilst the number of tweets related to COVID-19 children vaccination fluctuated in 2020, it steadily increased in 2021 with a distinct spike in the number of tweets in both May 2021 and October 2021. While the number of tweets gradually reduced in 2022, a marked increase was observed once again in June 2022.

month	1	2	3	4	5	6	7	8	9	10	11	12
year												
2020	0	41	275	542	653	357	725	610	585	2044	1737	2151
2021	1915	2045	3178	2922	7117	4354	5912	9311	13508	28178	15668	15275
2022	8682	6828	4800	3626	3508	6549	0	0	0	0	0	0

Table 1: Number of tweets per month - Details

5.1.2 Word level analysis

Word frequency

The fifty most commonly referenced words in the retrieved tweets are highlighted in

[7] Refer to Configuration Manual, sections 3.9 & 3.11

Figure 5. The first five words are linked to the themes of children, COVID-19 and vaccine while a clear dip occurs in the curve on the 6th most used word ('get').

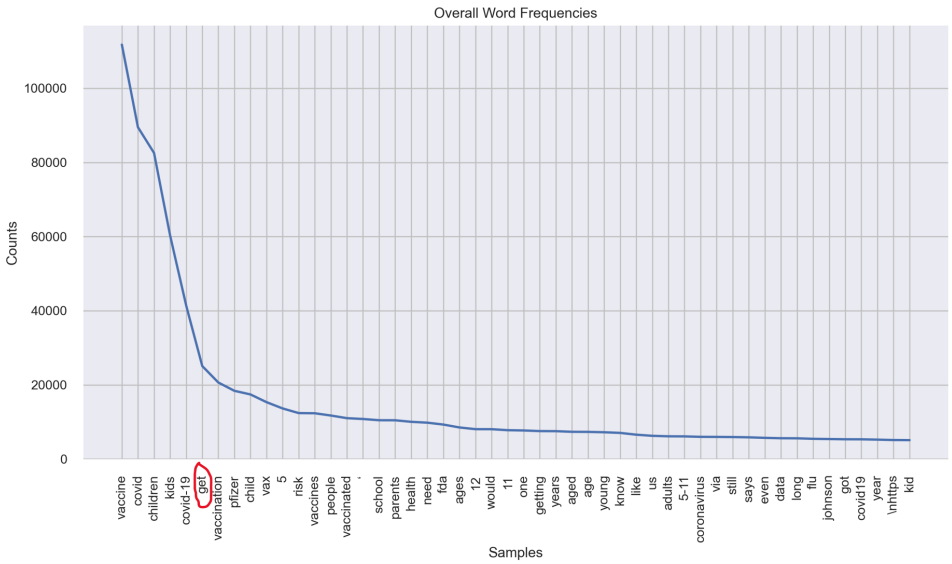

Figure 5: 50 most commonly used words

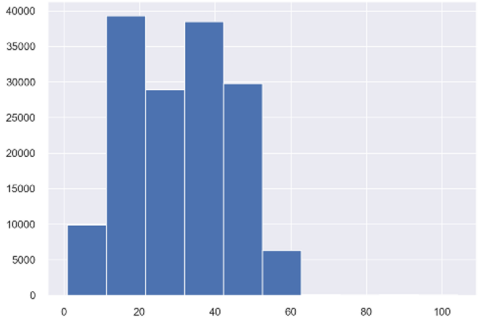

Figure 6: Overall word frequency

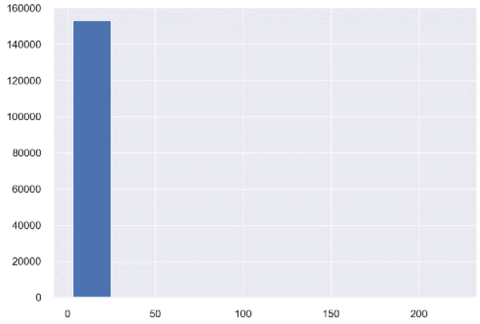

Figure 7: Average number of words in tweets

Multiple histograms were generated. Figure 6 shows that the words frequency in tweets range from 1 to 60 and is generally between 10 to 50 while the average number of words used range from 2 to 20 Figure 7. Furthermore, the average number of characters in tweets are in general between 100 and 280 [8].

Word clouds

Multiple word clouds were generated to give an overview of the most commonly referenced words (with a maximum of 100) in tweets, first globally as depicted in Figure 8 and then based on sentiment as shown in Figure 9.

The most frequent words in negative tweets are Johnson, unknown, repercussions, need, sick. Neutral tweets see high frequency of words such as COVID, Pfizer, paediatrician, health while positive tweets typically mention adjectives such as safe, effective, good, great, available, best, sure to describe vaccination amongst children.

[8]Refer to Configuration Manual, section 3.10

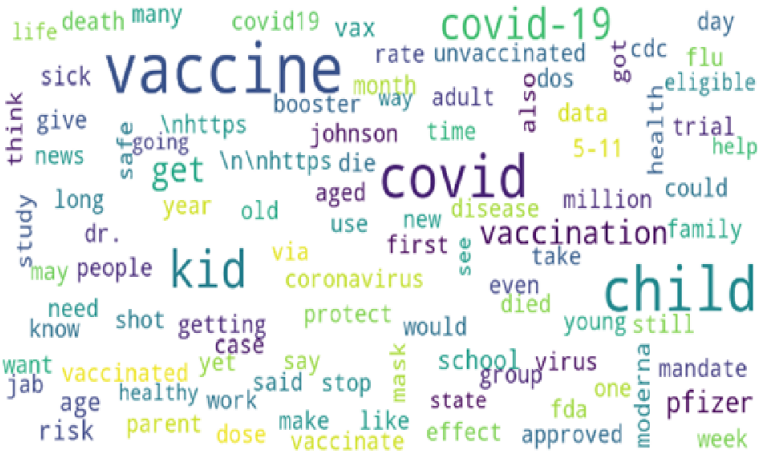

Figure 8: Overall Word Cloud

Figure 9: Word Cloud in negative, neutral and positive tweets

In addition, multiple word clouds of the most frequent words in negative, neutral and positive tweets mentioning specific vaccines manufacturers were also compiled[9].

N-Gram analysis

Insights into the text content of tweets were examined through the creation and analysis of bigrams, trigrams and quadgrams. Words combination were analysed, interpreted and potential topics identified in the corpus with Figure 10 illustrating the top ten quadgrams. The lexical field is quite negative with combinations of words like unknown, repercussions, down the, don't need, don't need the '**cking' clearly highlighting the fact that the authors are extremely sceptical of the need to vaccinate children as well as being concerned about potential unknown repercussions related to the COVID-19 vaccine in the future.

5.2 Emotion Analysis

Emotion and sentiment directed towards COVID-19 children vaccination programs was identified by using the National Research Council Canada (NRC) Word-Emotion Lexicon. The percentage of associated emotion-related words in the extracted tweets is depicted in Figure 11. The top three emotions identified were trust, anticipation and joy which all

[9]Refer to Configuration Manual, section 3.11

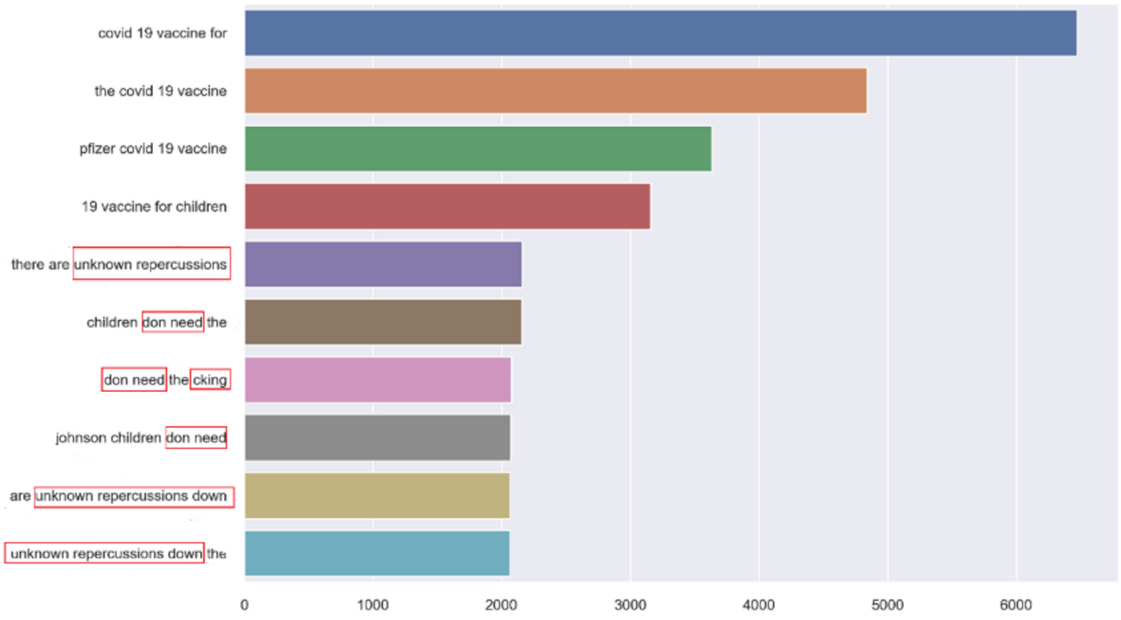

Figure 10: Top ten quadgrams

have a positive connotation. The percentage of negative emotion related words in tweets is smaller, with emotions in order of importance being fear, sadness and anger.

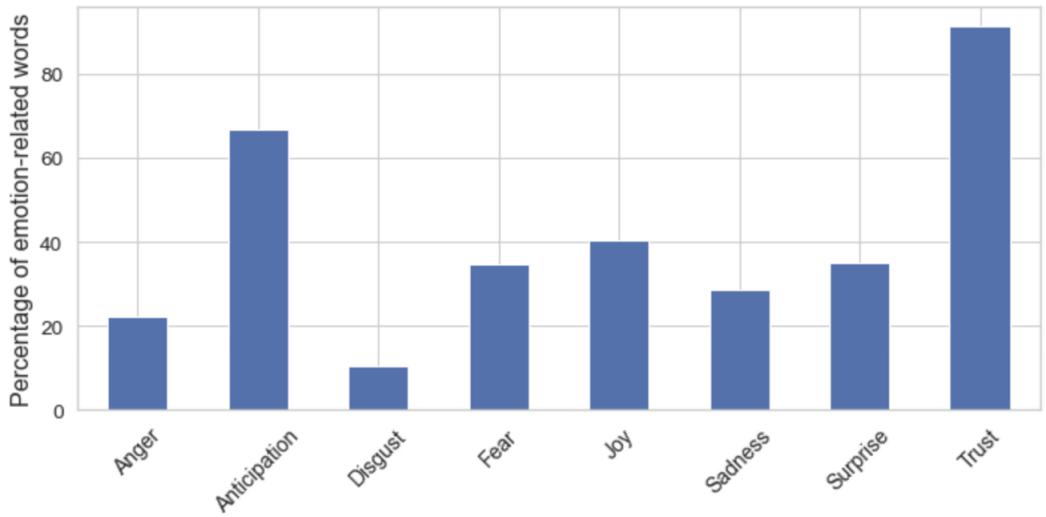

Figure 11: Percentage of emotion-related words in tweets

Furthermore, the top 10 words for 8 emotions were extracted and are represented in Figure 12. Emotions such as anger, disgust, fear and sadness use words such as shot, death, risk, infection while words such as young, safe, recommend, authorisation are present in the following emotions: anticipation, joy, surprise and trust.

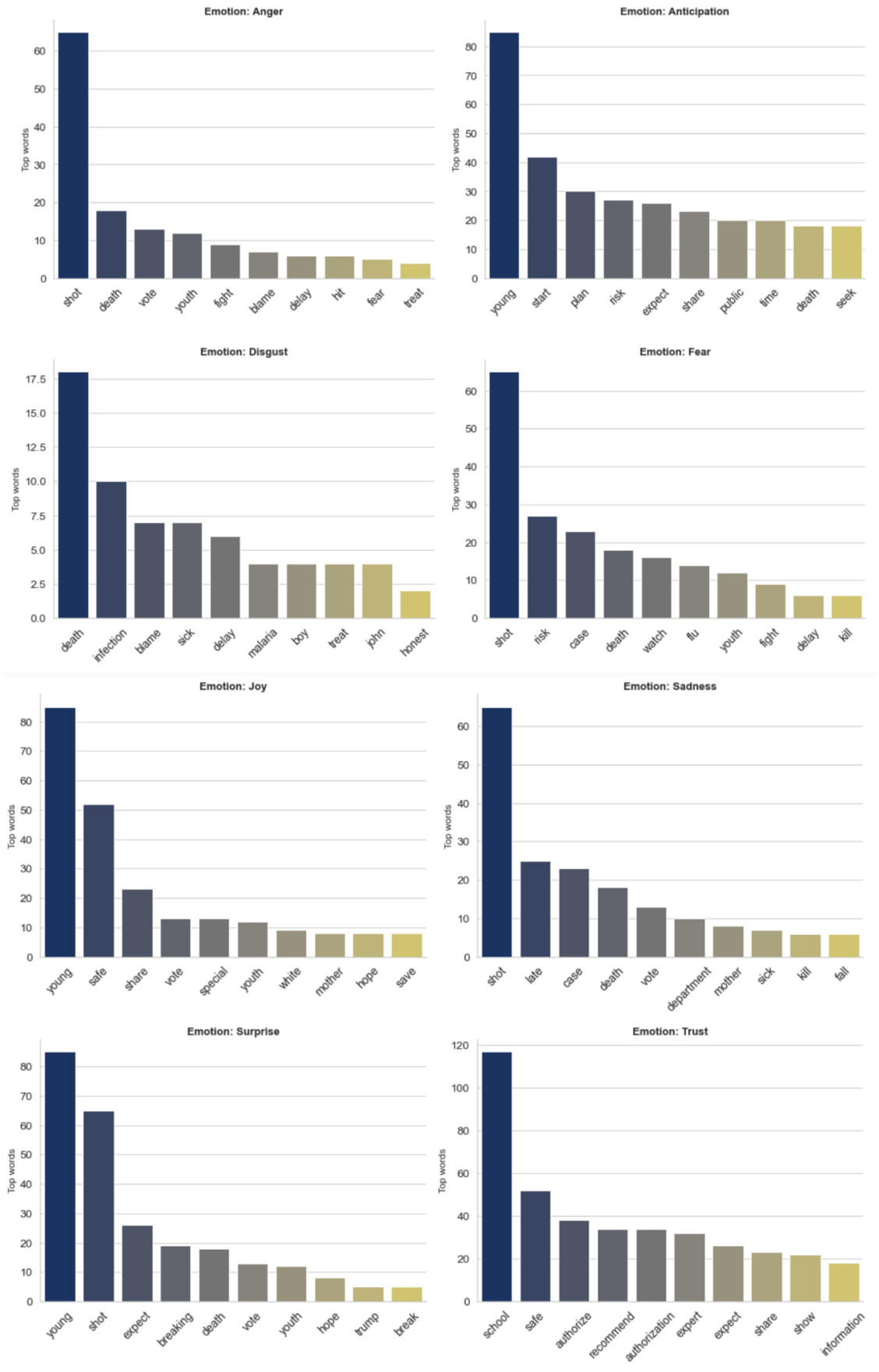

Figure 12: Top ten words for each emotion

5.3 Topic Modelling and Sentiment Analysis

5.3.1 Models implemented and their metrics

Multiple models were implemented in this research to analyse the data set's text-content so as to predict sentiment. Topics were modelled by using Latent Dirichlet Allocation modelling technique. The coherence score was used to assess the LDA model. Several supervised and deep learning models were developed and assessed to predict sentiment attached to the vaccination of children against COVID-19 such as logistic regression, random forest, extreme gradient boost, K-Nearest Neighbour, decision tree, naive bayes, simple RNN, single LSTM layer, bidirectional LSTM, 1D convolutional, BERT and DistilBERT. Furthermore, a regression analysis was carried out to predict sentiment attached to tweets based on the following predictors: number of cases, number of boosters administered and new vaccinations. The models used for this analysis were: logistic regression, random forest, extreme gradient boost, K-Nearest Neighbour, decision tree and support vector machine.

Various metrics were utilised to assess these models: accuracy, precision, recall, F1-Score, AUC and test loss.

5.3.2 Latent Dirichlet Allocation topic modelling

Latent Dirichlet Allocation (LDA) algorithm topic detection was applied to the tweets data set. The use of Python genism library allowed for the creation of dynamic graphs. Nine topics were identified in the tweets by the algorithm (as represented by bubbles in Figure 13).

The coherence score (0.37) was the highest with four topics. Words commonly used in the topic can be seen on the right of Figure 13. Words such as 'covid', 'vaccine' and 'risk' were widely used in topic 1 (31.4% of tokens) which discussed dangers attached to covid and vaccines while topic 8 (5.1% of tokens) highlights the use of more negatively charged words such as 'don't need', 'punish the unvaccinated', 'second grade citizens', 'wrong', 'unknown repercussions' or 'expose pharma' in tweets that question the use and need for vaccination against COVID-19 amongst children. The model returned some overlapping components (left-hand side of Figure 13) which demonstrate strong similarities between some of the topics extracted by the model.

Themes which were extracted and analysed are illustrated in Table 2 with strong similarities existing between Theme 1, 4 and 5, as well as between themes 2 and 3 (discussions specific to vaccines types and brands). Theme 8 and 9 clearly differ from the others with one being strongly negative (theme 8) whilst the other relatively positive (theme 9).

5.3.3 Sentiment analysis: polarity, subjectivity and objectivity in tweets

After an extensive text-content pre-processing and cleaning step, polarity, objectivity and subjectivity were calculated through the use of VADER and added in new columns in the tweets data set. These metrics are valuable tools to assess sentiment and the intensity linked to emotion. These scores were analysed through the development of multiple graphs to show the polarity and subjectivity scores related to tweets as depicted in Figure 14 and Figure 15 [10].

[10]Refer to Configuration Manual, section 3.11

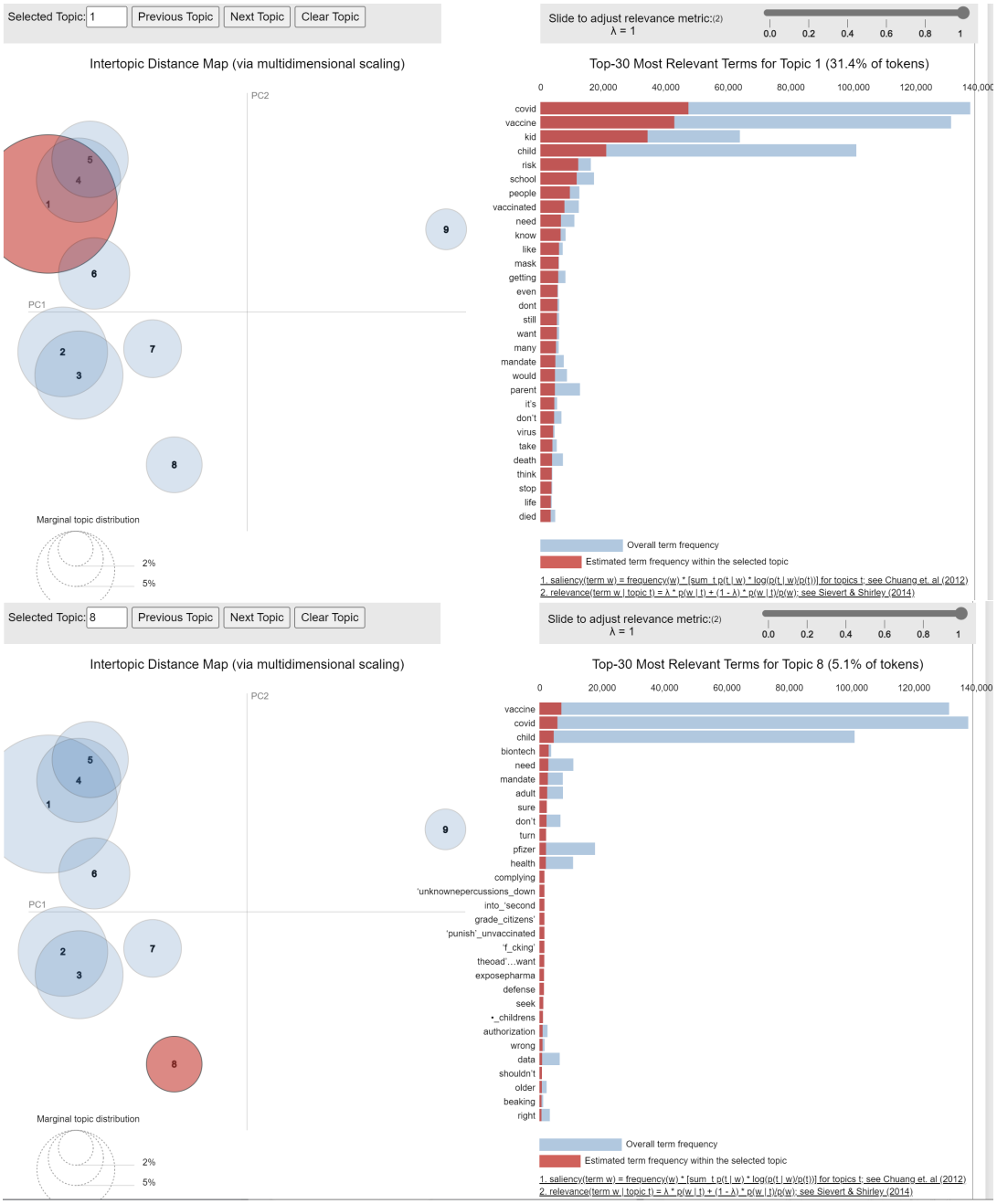

Figure 13: LDA topic detection - Dynamic graphs

Theme 1	Discussion on risks linked to the covid-19 virus and vaccines faced by children
Theme 2	Discussion on vaccines effectiveness (Pfizer and Moderna), availability, number of doses required and existing studies
Theme 3	Discussion on Pfizer and Moderna vaccines approval, appointments and clinics
Theme 4	Discussion on hospital, healthcare and deaths
Theme 5	Discussion on how getting the entire family vaccinated is best
Theme 6	Discussion on data on vaccines, efficacy, cases, myocarditis and situation in the state of Florida
Theme 7	Discussion on how paediatricians should be questioned and how parents should get full answers
Theme 8	Questions the use and need of children vaccination against COVID-19. Discussion on Pfizer vaccine, the unknown repercussions of the vaccine, criticism of the pharmaceutical companies, mention of second-grade citizens and how the unvaccinated are punished
Theme 9	Vaccines are safe and important to protect the vulnerable

Table 2: Detected topics

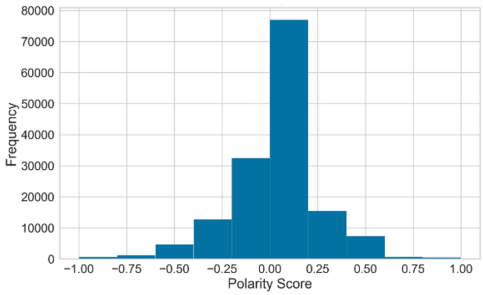

Figure 14: Polarity

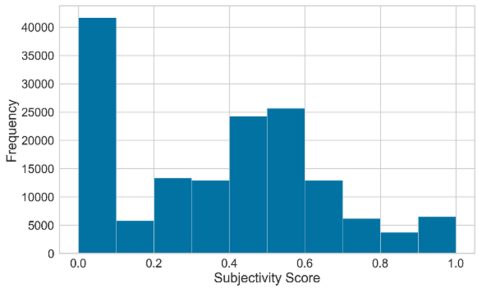

Figure 15: Subjectivity

In addition, graphs depicting the average subjectivity and polarity scores for overall tweets related to the COVID-19 vaccine (Figure 16) as well as tweets specific to vaccine manufacturers[11] were generated to further analyse the data.

Sentiment attached to text-content was extracted using the polarity scores. These scores were used to categorise tweets into three classes (negative with polarity between -1 and -0.01, neutral with polarity between -0.01 and 0.01, positive with polarity between 0.01 and 1).

5.3.4 Sentiment prediction modelling

A separate csv file was generated so as to facilitate the implementation of sentiment prediction models. It is a simplified version of the original tweets data set used during the exploratory phase which holds only the columns required for the analysis such as the date, text, sentiment, negative score, positive score, neutral score and compound score. Numerous supervised models were researched, implemented and compared: logistic regression, random forest, extreme gradient boost, K-Neighbours, decision tree, naive bayes. The following deep learning models were developed as well: simple RNN, single LSTM layer model, Bidirectional LSTM, 1D convolutional, Bi-directional Encoding Representation for a Transformer (BERT) and DistilBERT. Multiple tests were subsequently executed to fine-tune these models, for instance through multiple iterations of different epochs, with associated tests results outlined in the next section.

[11]Refer to Configuration Manual, section 3.11

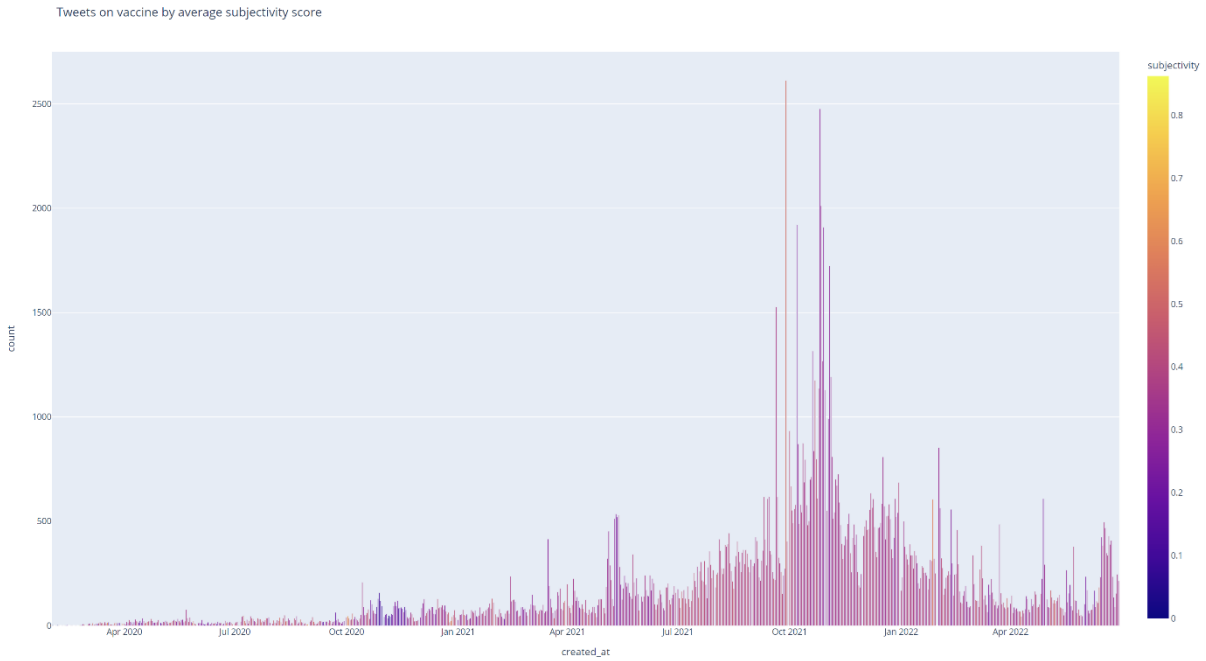

Figure 16: Tweets on vaccine by average subjectivity

6 Evaluation and Results

Polarity, subjectivity and objectivity in tweets

The sentiment polarity within the time-frame of February 2020-June 2022 has been captured as part of this analysis and is displayed in Figure 17. Neutral emotions are mostly prevalent in tweets throughout the period as demonstrated in Cotfas et al. (2021) however sentiment is predominantly negative between July-September 2020, the period three months prior to the authorisation in the U.S. of the Pfizer-BioNTech vaccine amongst children in the 5-11 age bracket.

Tweets' subjectivity and objectivity scores are depicted in Figure 18 where sentiment is mostly objective. The portion of tweets classified as subjective was very similar to those marked as objective in September 2021 but this tendency shifted greatly in October 2021 with tweets being overwhelmingly objective. This trend is repeated in tweets posted between May 2021 and June 2022.

The polarity to subjectivity is shown in Figure 19 where it appears that neutral tweets are overwhelmingly objective. Negative tweets are slightly more subjective while positive ones are more objective than subjective.

Text-Content sentiment prediction

Table 3 lists the classical models which were implemented and their associated results. Random forest and decision tree models were the best performing models in terms of accuracy, precision, recall and F1-Score.

Six types of deep learning models were implemented and fine-tuned. Multiple tests were carried out using bi-directional LSTM, single LSTM, 1D Convolutional and simple RNN models with epoch set to the following values: 70, 200, 80, 90 and 85. Bidirectional LSTM model constantly outperformed the other ones in this series of tests, with an accuracy close to 90% with epoch set to 85. In addition, BERT and DistilBERT models

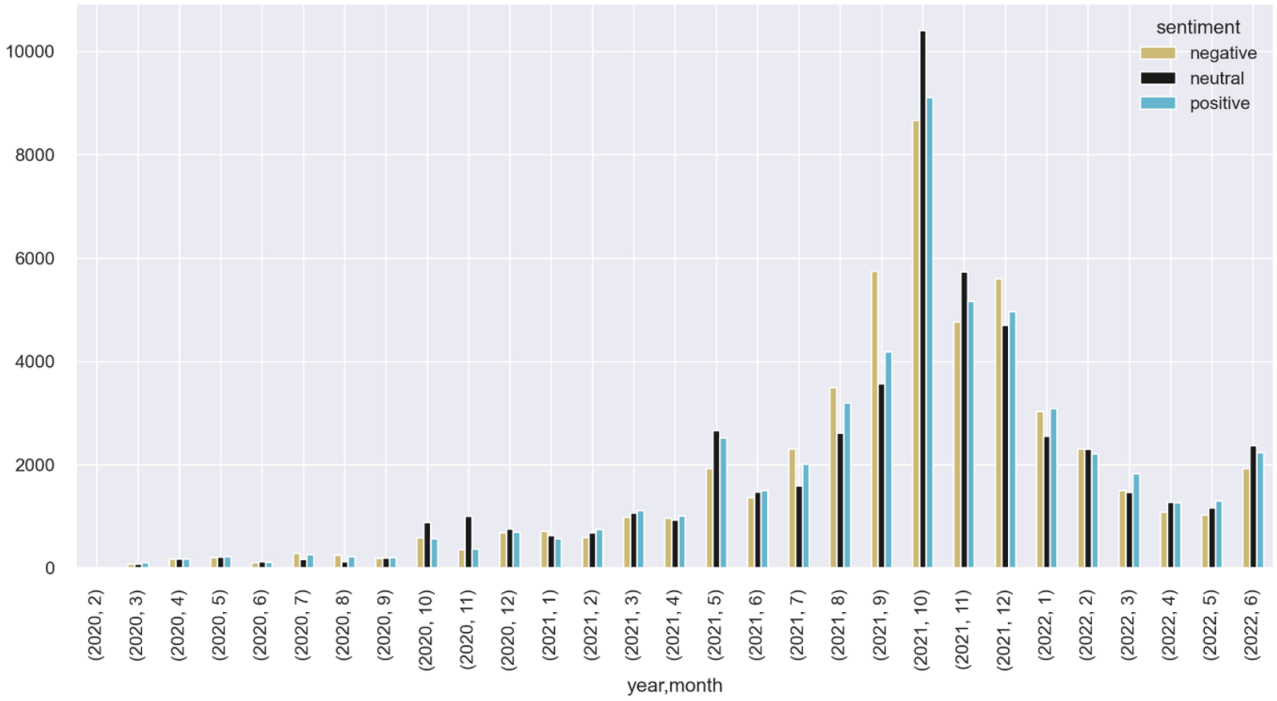

Figure 17: Sentiment polarity

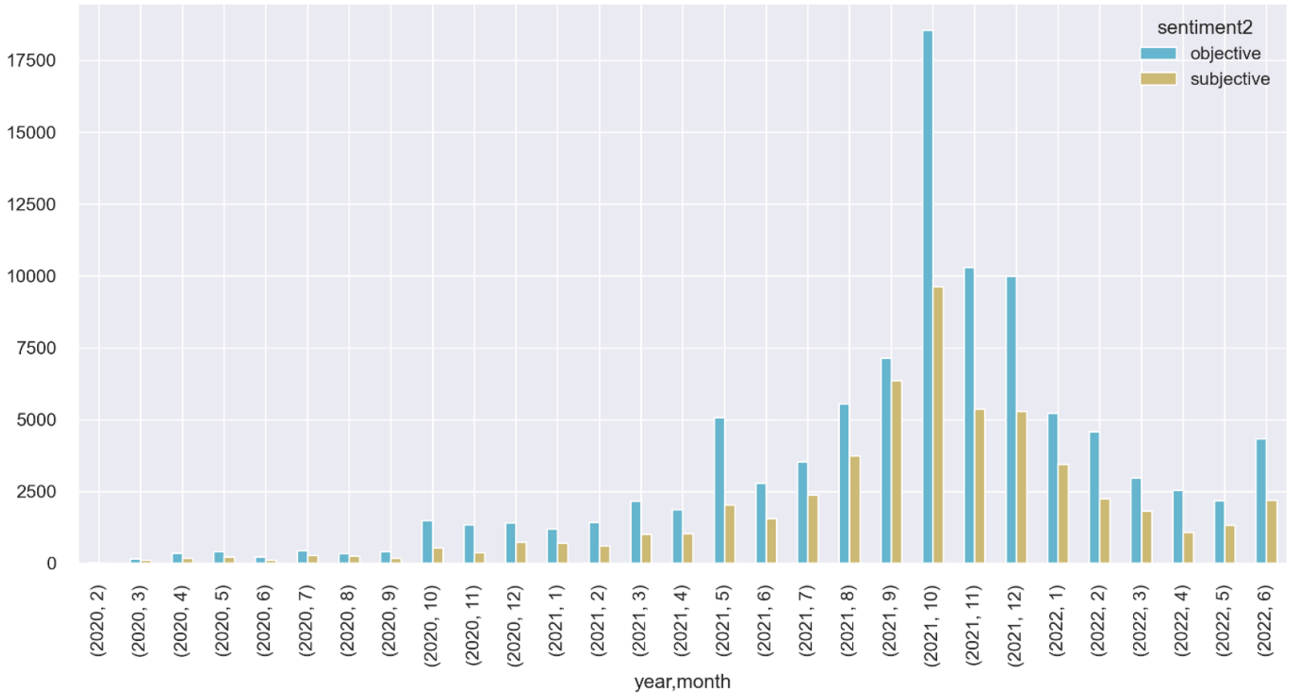

Figure 18: Objectivity and subjectivity in tweets

were developed and ran with epoch set to 2. BERT was the best performing model overall with a test accuracy of over 90% and loss of 30% as shown in Table 4.

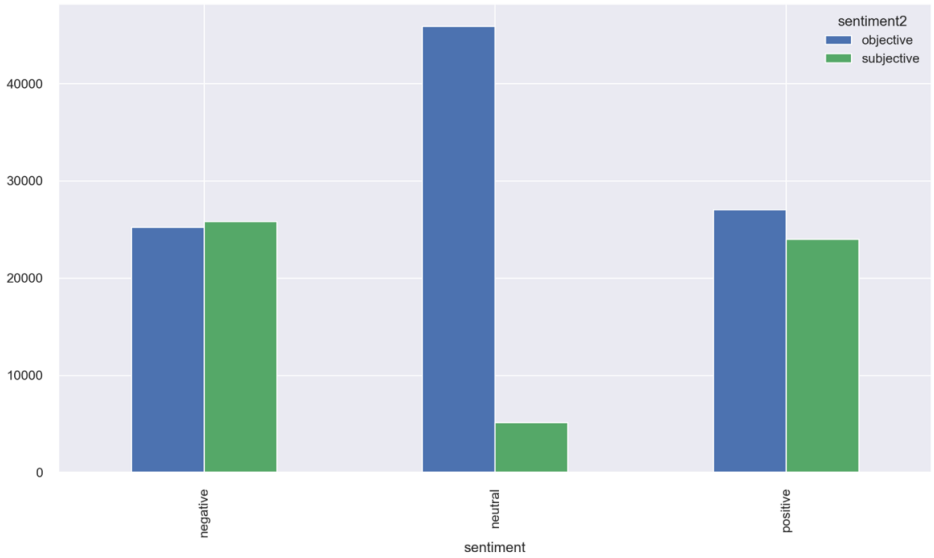

Figure 19: Polarity to subjectivity

Model	Accuracy	Precision	Recall	F1-Score
Random Forest (Train)	0.70	0.70	0.70	0.69
Random Forest (Test)	0.85	0.86	0.85	0.85
Decision Tree (Train)	0.67	0.68	0.67	0.66
Decision Tree (Test)	0.82	0.83	0.82	0.82
Support Vector (Train)	0.59	0.59	0.59	0.58
Support Vector (Test)	0.65	0.65	0.66	0.65
K-Nearest Neighbour (Train)	0.56	0.56	0.56	0.55
K-Nearest Neighbour (Test)	0.63	0.64	0.63	0.62
Extreme Gradient Boost (Train)	0.57	0.57	0.57	0.56
Extreme Gradient Boost (Test)	0.61	0.63	0.61	0.61
Logistic Regression (Train)	0.52	0.51	0.52	0.50
Logistic Regression (Test)	0.52	0.51	0.52	0.50

Table 3: Supervised models - Results comparison

Model	Epoch	Accuracy
BERT	2	90.3%
DistilBERT	2	89.84%
Bidirectional LSTM	85	89.94%
Single LSTM	85	89.11%
1D Convolutional	85	82.06%
Simple RNN	85	78.23%

Table 4: Deep Learning models - Results comparison

Sentiment prediction with combined data sets

The sentiment analysis carried out in this research was correlated with data extracted

from 'Our World in Data'[12] website to highlight and monitor whether emotions and sentiment attached to the vaccination of children against COVID-19 evolved as the rollout continued and the number of cases fluctuated. The research undertaken highlighted that the model with the highest accuracy was Random Forest with an accuracy of 41.63% [13]..

7 Discussion

Volume analysis
Analysis of the frequency of tweets posted on a monthly basis showed a distinct correlation between the number of tweets posted and on-going decisions and announcements emerging from public health organisations and decision makers. For example there was a noticeable increase in the number of tweets in November 2020 when Pfizer-Biontech announced they had developed a vaccine which had a 90% effective rate on November 9, 2020. Similarly, there was a surge in the amount of tweets relating to children vaccinated against COVID-19 in May 2021 when the emergency approval for the use of the Pfizer-BioNTech vaccine was extended by the U.S. Food and Drug Administration (FDA) to include children between the age of 12 and 15. The highest recording of tweets referencing children vaccinations and COVID-19 occurred in October 2021 when the U.S. FDA authorised the vaccination of children aged 5-11 with the Pfizer-BioNTech vaccine. This trend continued the following month when on November 2nd, 2021 the Center for Disease Control and prevention (CDC) expanded vaccine recommendations to approximately 28 million children in the U.S. in this 5-11 year old age cohort and strongly urged vaccination programs to commence as soon as possible. There was a marked reduction in the number of tweets between April and May 2022, however, the number of tweets increased yet again in June 2022 as health officials cleared the use of the Moderna and Pfizer-BioNTech coronavirus vaccines for very young children (aged between 6 months and 5 years old) in the USA on June 18th, 2022. This temporal analysis illustrated that by the end of December 2021, tweets were more negative than positive or neutral whilst the findings suggested that the public's perception of vaccines shifted and that neutral/positive sentiment had faded over time. From this research, it can be confirmed that doubts and unfounded concerns need to be further addressed by public policy makers through the development of continuous, strengthened and more focused vaccination education programs after the authorisation of vaccination rollouts amongst particular age groups of the population.

Number of Tweets posted by official news outlets
The frequency of tweets posted by a number of official paper/news channels between February 2020 and June 2022 was also analysed with results outlined in Table 5. Given the relatively low volume of tweets posted by such media outlets, it can be concluded that Twitter is a social media platform that has the potential to be used more extensively by news agencies to inform and better educate its subscribers about COVID-19 children vaccination programs.

[12]Our World in Data: `https://ourworldindata.org/coronavirus`
[13]Refer to Configuration Manual, section 3.17

Name	Twitter Account	Number of Tweets
Reuters	Reuters	65
New York Times	nytimes	34
The Independent	Independent	33
CNN	CNN	25
US News	usnews	23
Washington Post	washingtonpost	19
Time	TIME	15
Newsweek	Newsweek	15
MSNBC	MSNBC	14
Guardian	guardian	12
Guardian News	guardiannews	9
The Sun	TheSun	8
The Times	thetimes	8
New York Post	nypost	7
Irish Times	IrishTimes	7
France 24	France24$_en$	6
CNN breaking	cnnbrk	6
bbcnews	BBCNews	6
The Mirror	DailyMiror	5
huffingtonpost	HuffPost	4
New Scientist	newscientist	3
Politico	politico	3
Fox News	FoxNews	3
RTE Series	rte	0
The Economist	theeconomist	0

Table 5: Number of tweets - Official news outlets

Correlation with 'Our World in Data' statistics

With respect to the analysis of data retrieved form the 'Our World in Data' website, metrics such as accuracy, precision, recall and f1-score output by the various models were compared. Despite the fact that the data sample size was sufficiently adequate (153,096), the best performing model (random forest) had an accuracy of just above 40%. Such results can be explained by a very weak correlation between sentiment and the metrics selected in this second data set. As this research focused on tweets posted at a world-wide level, some of the statistics in the second data set could not be exploited. Whilst filtering tweets based on specific geographical locations would have resulted in smaller data samples size, it could however have allowed for the analysis and comparison of sentiment against predictors such as population median age, aged over 65 or 70, GDP per capita, life expectancy, cardiovascular death rate, diabetes prevalence, female smokers, male smokers, handwashing facilities, hospital beds per thousand and human development index. It could be argued that findings from such analysis would have proved to be most insightful.

Modelling

The results output from this research have been compared against previous analysis as

depicted in Table 6 and Table 7 with the classical models broadly in line with previous findings. While the random forest and naive bayes models developed in this study performed better than some existing research, this was not the case for support vector machine and extreme gradient boost algorithms. Deep learning models implemented in this study outperformed the selected existing research's results such as Bidirectional LSTM and Single LSTM with accuracies very close to 90%. Similarly, results obtained while executing BERT and DistilBERT models produced high accuracy of 90% and over which is promising and in line with existing research.

Research	**Model**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
This paper	RF	0.85	0.86	0.85	0.85
Paul and Gokhale (2021)	RF	0.87	0.92	0.68	0.78
Nurdeni et al. (2021)	RF	0.75	0.77	0.75	0.70
This paper	Extreme GB	0.61	0.63	0.61	0.61
Paul and Gokhale (2021)	Extreme GB	0.84	0.80	0.68	0.78
This paper	SVM	0.65	0.65	0.66	0.65
To et al. (2021)	SVM	0.92	0.20	0.75	0.32
Paul and Gokhale (2021)	SVM	0.86	0.87	0.68	0.74
Du et al. (2017)	SVM	0.78	0.78	0.78	0.78
Nurdeni et al. (2021)	SVM	0.77	0.76	0.77	0.75
This paper	NB	0.70	0.70	0.70	0.70
Nurdeni et al. (2021)	NB	0.69	0.61	0.69	0.59
To et al. (2021)	NB	0.88	0.23	0.32	0.27

Table 6: Supervised models results - Comparison with existing research

Research	**Model**	**Accuracy**
This paper	BERT	90.3%
K.Hayawi et al. (2022)	BERT	98%
Müller et al. (2020)	BERT	93.1%
To et al. (2021)	BERT	92.5%
Aygun et al. (2022)	BERT	86%
Cotfas et al. (2021)	BERT	78.94%
This paper	Bidirectional LSTM	89.94%
To et al. (2021)	Bidirectional LSTM	88.4%
Cotfas et al. (2021)	Bidirectional LSTM	74.74%
This paper	Single LSTM	89%
Paul and Gokhale (2021)	Single LSTM	82%

Table 7: Deep Learning models results - Comparison with existing research

However, there is scope to further fine-tune the models by increasing the number of epochs while running these algorithms. Similarly, this research results indicate that effective opinion mining of children COVID-19 vaccination programs can be an efficient tool to assess social media users' sentiment which in turn could be used by health officials and policy-makers to design dynamic policies to reach decisions in a timely manner. This information could also be used to create more detailed, targeted vaccine campaigns.

Although determining the originating geographical location of Twitter posts was a challenge, it was decided that the focus of this research would be directed more towards global sentiment as the corpus size was of greater importance. As the source data used in this research are posts written in English extracted from a single social media platform (Twitter), it is acknowledged that this data may not be representative of the worldwide population as external factors such as socio-demographics greatly influence the type of users of a given social media platform. This research initiative could potentially have benefited from using enriched data sets as this would help improve the understanding of public sentiment attached to the vaccination of children against COVID-19 as well as any underlying behavioural determinants. This could be achieved by extending the keywords used to scrape data from Twitter as well as using data extracted from other social media platforms. Finally, while numerous models were implemented in this particular research, analysis could have been further augmented through the investigation and potential inclusion of unsupervised deep neural networks (Brandt et al. (2020), Mahmud et al. (2018)) and fuzzy-based approaches (Hussain et al. (2021)).

8 Conclusion and Future Work

COVID-19 is the first pandemic of its kind to occur in the age of social media. Gaining a good understanding of emotions and sentiment attached to children vaccination programs is crucial for effective policy making. This research involved data being scraped from the Twitter platform before being analysed through the use of multiple artificial intelligence techniques. Combining sentiment analysis with topics detection allowed for trends and discussions about COVID-19 children vaccination to be highlighted and concerns identified. An analysis of public emotions showed that the top three emotions were trust, anticipation and joy while the percentage of negative emotion related words in tweets was smaller, with emotions expressed (in order of importance) being fear, sadness and anger. Results from this research provided insight into some of the reasons behind public hesitancy towards COVID-19 children vaccination programs as recurrent themes in negative tweets were that children 'don't need' (to be vaccinated), that the unvaccinated population were being 'punished' and treated as 'second-grade citizens' and that the vaccination of children was 'wrong' and could have 'unknown repercussions'. This research also illustrated that tweets between February 2020 and June 2022 were mainly neutral. Sentiment was more positive than negative, apart from July-September 2021 however, which is the time-frame leading up to the authorisation of the vaccination of children against COVID-19 in the USA which occurred in October 2021. Furthermore, tweets were more objective than subjective, especially in May 2021, October 2021 and June 2022 where decisions regarding the vaccination of children against COVID-19 were announced in the US. The models developed offered valuable insights on topics and sentiment linked to children vaccination. A polarity to subjectivity analysis showed that neutral tweets were overwhelmingly objective, with negative tweets being slightly more subjective than objective whilst positive posts were more objective than subjective. After implementing and comparing 12 classical and deep learning methods, the research highlighted that the best performing model was BERT with an accuracy of 90.3%. After further fine-tuning, results produced by models such as BERT or DistilBERT could support policy-makers understand better population's concerns, opinion and sentiment. This information could be used by authorities and health organisations to enhance communic-

ation strategies to inform, reassure and fight unproven concerns amongst the population to increase worldwide vaccination rollouts.

While current findings could be used by public health organisations for the development of more effective, tailored health education and vaccination campaigns, this research could be extended in multiple ways. Firstly, the text-content analysis focused solely on data scraped from Twitter. Investigating data extracted from multiple social and news media platforms would further enrich the text-content in the analysis of emotions and themes attached to COVID-19 children vaccination campaigns. Whether or not there are differences in the terms and rhetoric used in official news outlets based on political influence would be an interesting angle to investigate. Another area of interest might be the analysis of social media posts based on authors' location and highlight any influence external factors such as GDP, life expectancy or median population age may have on posters' sentiment. Future studies could also consider analysing the number of followers and friends of posters and investigating whether these numbers have any influence on the text-content's emotion and sentiment would be valuable. Furthermore, analysing hashtags used and any correlation existing between these hashtags and the number of likes and re-tweets a given tweet received would be another interesting study. The research could be further enriched by analysing tweets that include URLs and check whether the sentiment attached to such tweets is mostly neutral. Similarly, analysing verified Twitter authors' sentiment towards COVID-19 children vaccination and comparing it against unverified ones would also be insightful. Finally, running the developed models on data sets specific to COVID-19 vaccines authorised for the inoculation of children (i.e. Pfizer and Moderna) could also offer more granular results specific to these two vaccines which could be used by authorities and manufacturers to adjust and tailor their communication strategies appropriately.

9 Acknowledgments

I would like to thank my supervisor Dr. Vladimir Milosavljevic for all his support and guidance throughout the implementation of this research project.

References

Abd-Alrazaq, A., Alhuwail, D., M., M. H. H. and Shah, Z. (2020). Top concerns of tweeters during the covid-19 pandemic: A surveillance study.

Alam, K. N., Khan, M. S., Dhruba, A. R., Khan, M. M., Al-Amri, J. F., Masud, M. and Rawashdeh, M. (2021). Deep learning-based sentiment analysis of covid-19 vaccination responses from twitter data.

Aygun, I., Kaya, B. and Kaya, M. (2022). Aspect based twitter sentiment analysis on vaccination and vaccine types in covid-19 pandemic with deep learning.

Boucher, J.-C., Cornelson, K., Benham, J. L., Fullerton, M. M., Tang, T., Constantinescu, C., Mourali, M., Oxoby, R. J., Marshall, D. A., Hemmati, H., Badami, A., Hu, J. and Lang, R. (2021). Analyzing social media to explore the attitudes and behaviors following the announcement of successful covid-19 vaccine trials: Infodemiology study.

Brandt, J., Buckingham, K., Buntain, C., Anderson, W., Ray, S., Pool, J.-R. and Ferrari, N. (2020). Identifying social media user demographics and topic diversity with computational social science: a case study of a major international policy forum.

Cinelli, M., Quattrociocchi, W., AlessandroGaleazzi, MicheleValensise, C., Brugnoli, E., Schmidt, A. L., Zola, P., Zollo, F. and Scala, A. (2021). The covid-19 social media infodemic.

Cotfas, L.-A., Delcea, C., Roxin, I., Ioanăş, C. and Gherai, D. S. (2021). The longest month: Analyzing covid-19 vaccination opinions dynamics from tweets in the month following the first vaccine announcement.

Du, J., Xu, J., Song, H.-Y. and Tao, C. (2017). Leveraging machine learning-based approaches to assess human papillomavirus vaccination sentiment trends with twitter data.

Dunn, A. G., Leask, J., Zhou, X., Mandl, K. D. and Coiera, E. (2015). Associations between exposure to and expression of negative opinions about human papillomavirus vaccines on social media: An observational study.

Evanega, S., Lynas, M., Adams, J. and Smolenyak, K. (2020). Coronavirus misinformation: quantifying sources and themes in the covid-19 'infodemic'.

Garcia, K. and Berton, L. (2021). Topic detection and sentiment analysis in twitter content related to covid-19 from brazil and the usa.

González-Padillal, D. A. and Tortolero-Blanco, L. (2020). Social media influence in the covid-19 pandemic.

Hussain, A., Tahir, A., Hussain, Z., Sheikh, Z., Gogate, M., Dashtipour, K., Ali, A. and Sheikh, A. (2021). Artificial intelligence–enabled analysis of public attitudes on facebook and twitter toward covid-19 vaccines in the united kingdom and the united states: Observational study.

Kaur, S., Kaul, P. and Zadeh, P. M. (2020). Monitoring the dynamics of emotions during covid-19 using twitter data.

K.Hayawi, S.Shahriar, M.A.Serhani, I.Taleba and S.S.Mathewa (2022). Anti-vax: a novel twitter dataset for covid-19 vaccine misinformation detection.

Kornides, M. L., Badlis, S., Head, K. J., Putt, M., Cappella, J. and Gonzalez-Hernadez, G. (2022). Exploring content of misinformation about hpv vaccine on twitter.

Krawczyk, K., Chelkowski, T., Laydon, D. J., Mishra, S., Xifara, D., Gibert, B., Flaxman, S., Mellan, T., Schwämmle, V., Röttger, R., Hadsund, J. T. and Bhatt, S. (2021). Quantifying online news media coverage of the covid-19 pandemic: Text mining study and resource.

Luo, X., Zimet, G. and Shah, S. (2019). A natural language processing framework to analyse the opinions on hpv vaccination reflected in twitter over 10 years (2008 - 2017).

Lyu, J. C., Han, E. L. and Luli, G. K. (2021). Covid-19 vaccine–related discussion on twitter: Topic modeling and sentiment analysis.

Mahmud, M., Kaiser, M. S., Hussain, A. and Vassanelli, S. (2018). Applications of deep learning and reinforcement learning to biological data.

Martin, S., Kilich, E., Dada, S., Kummervold, P. E., Denny, C., Paterson, P. and Larson, H. J. (2020). "vaccines for pregnant women...?! absurd" – mapping maternal vaccination discourse and stance on social media over six months.

Merchant, R. M. and Lurie, N. (2020). Social media and emergency preparedness in response to novel coronavirus.

Müller, M., Salathé, M. and Kummervold, P. E. (2020). Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter.

Nurdeni, D. A., Budi, I. and Santoso, A. B. (2021). Sentiment analysis on covid19 vaccines in indonesia: From the perspective of sinovac and pfizer.

Paul, N. and Gokhale, S. S. (2021). Analysis and classification of vaccine dialogue in the coronavirus era.

Piltch-Loeb, R., Savoia, E., Goldberg, B., Hughes, B., Verhey, T., Kayyem, J., Miller-Idriss, C. and Testa, M. (2021). Examining the effect of information channel on covid-19 vaccine acceptance.

Rahul, K., Jindal, B. R., Singh, K. and Meel, P. (2021). Analysing public sentiments regarding covid-19 vaccine on twitter.

Reddy, D. M. and Reddy, N. V. S. (2019). Twitter sentiment analysis using distributed word and sentence representation.

Ruiz, J. and Bell, R. (2021). Predictors of intention to vaccinate against covid-19: Results of a nationwide survey.

Shah, Z., Surian, D., Dyda, A., Coiera, E., Mandl, K. D. and Dunn, A. G. (2019). Automatically appraising the credibility of vaccine-related web pages shared on social media: A twitter surveillance study.

Shamrat, F., S., C., M.M., I., J.N., M., Md.M., B., P., D. and Md.O., R. (2021). Sentiment analysis on twitter tweets about covid-19 vaccines using nlp and supervised knn classification algorithm.

Shapiro, G. K., Surian, D., Dunn, A. G., Perry, R. and Kelaher, M. (2017). Comparing human papillomavirus vaccine concerns on twitter: a cross-sectional study of users in australia, canada and the uk.

Tang, L., Bie, B., Park, S.-E. and Zhi, D. (2018). Social media and outbreaks of emerging infectious diseases: A systematic review of literature.

To, Q. G., To, K. G., Huynh, V.-A. N., Nguyen, N. T. Q., Ngo, D. T. N., 1, S. J. A., Tran, A. N. Q., Tran, A. N. P., Pham, N. T. T., Bui, T. X. and Vandelanotte, C. (2021). Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter.

Xue, J., Chen, J., Hu, R., Chen, C., Zheng, C., Su, Y. and Zhu, T. (2021). Twitter discussions and emotions about the covid-19 pandemic: Machine learning approach.