Beyond the Tweet: Exploring Reddit as an Alternative Resource for Disease Forecasting

Nathanael Bashford

28 February 2024

Independent Research Project: COM00151M

Supervisor: Waseem Ahmed

**ACKNOWLEDGEMENTS**

I would like to thank my project supervisor, Waseem Ahmad, for all the help, guidance, and patience he has given to me throughout this project.

I would also like to thank my family and friends for their support throughout this journey, and my wife, for her constant love and reassurance.

**EXECUTIVE SUMMARY**

Disease forecasting, a crucial component of the strategic response to pandemics like Covid-19, increasingly relies on digital data, offering large-scale immediate insights, ultimately helping to save lives [1]. However, Twitters recent paywall restrictions limits this crucial resource [2], requiring exploration of alternative resources to maintain preparedness against future pandemics.

**Aims and Objectives**

Our research aims to evaluate Reddits unexplored utility for disease forecasting, employing a case study design, with a longitudinal focus on the initial Covid-19 pandemic period between 1st March – 31st December 2020 in the USA. Our objectives centre on assessing Reddits feasibility for forecasting, how suitable methodologies can be adopted, and to empirically measure Reddits forecasting utility.

**Methods**

Covid-19 data was collected from the World Health Organisation, and Reddit data from an opportunistic dataset from SocialGrep.com. For information extraction, we classify Reddit texts along an ‘Author’ and ‘Contact’ dimension, capturing direct and indirect accounts of Covid-19 infections respectively. We train BERT models to classify Reddit texts, then acquire time-series data summarising temporal changes in Reddit classification label frequency.

We simulate a real-world forecasting exercise, generating overlapping training and test sets using a Rolling-Window approach, then predict Covid-19 incidence 1, 7 and 14-days ahead for both baseline and Reddit-enhanced Random Forest Regression (RFR) models using a Walk-forward testing approach.

**Findings**

Our exploration revealed Reddit provides direct and indirect indicators of Covid-19 incidence and trajectory. Our Reddit-enhanced RFR models improved baseline prediction errors (MAE) by 23.26% and 31.83% for 7 and 14-day forecasts respectively and reduced the MAPE variability to 4.74% from the baseline's 10.18% at 14-day forecasts; indicating our Reddit features produce more robust and accurate Covid-19 forecasting models, particularly at longer forecast horizons.

Of our Reddit features, personal accounts of infection were consistently the strongest predictor to Covid-19 incidence, with a strong correlation of 0.95, whilst suspected/ambiguous and second-hand accounts showed weaker correlations, 0.37 and 0.70 respectively, and less informative for predictions.

**Limitations**

However, no improvement was observed for immediate 1-day ahead predictions, and Reddit features provided little information at a significant Covid-19 peak, questioning predictive reliability.

Our exploration likewise revealed inherent challenges in Reddits real-world application, as its lengthy texts and lack of geospatial data limits both timely and specific insights, critical requirements for disease surveillance [3].

**Ethics Considerations**

Though Reddit users post content in the public sphere, issues of informed consent arise since users are unaware their posts are utilised for insights and real-world applications. This likewise links to privacy issues, as textual exploration may inadvertently identify individuals who have a right to anonymity. Whilst disease forecasting aims to benefit public health, underrepresented users and communities within forecasting data may produce biased conclusions, potentially harming vulnerable populations [4].

**Conclusion**

Despite these challenges, our exploration revealed Reddit provides comparable classification-based indicators to Covid-19 as Twitter, highlighting the importance of investigating alternative digital resources and improving collective preparedness for future pandemics.

Table of Contents

[List of Figures 6](#_Toc160010447)

[List of Tables 7](#_Toc160010448)

[1 Introduction 8](#_Toc160010449)

[1.1 Disease Forecasting 9](#_Toc160010450)

[1.2 Digital Disease Forecasting 9](#_Toc160010451)

[1.3 Reddit for Forecasting 10](#_Toc160010452)

[1.4 Aims and Objectives 11](#_Toc160010453)

[1.5 Research Focus 12](#_Toc160010454)

[1.6 Report Structure 12](#_Toc160010455)

[2 Literature Review 14](#_Toc160010456)

[2.1 Information-Extraction 15](#_Toc160010457)

[2.2 Optimal Classification technique 17](#_Toc160010458)

[2.3 Forecasting techniques 19](#_Toc160010459)

[2.4 Research Questions 21](#_Toc160010460)

[3 Methodology 23](#_Toc160010461)

[3.1 Research Philosophy 23](#_Toc160010462)

[3.2 Methodology Overview 24](#_Toc160010463)

[3.3 Data Sources 25](#_Toc160010464)

[3.3.1 Covid-19 25](#_Toc160010465)

[3.3.2 Reddit Data 25](#_Toc160010466)

[3.3.3 Pre-Processing 26](#_Toc160010467)

[3.4 Reddit Information Extraction 27](#_Toc160010468)

[3.5 Text Classification 29](#_Toc160010469)

[3.5.1 BERT 29](#_Toc160010470)

[3.5.2 Classification Evaluation 30](#_Toc160010471)

[3.5.3 BERT Validation and Testing 31](#_Toc160010472)

[3.6 Forecasting 32](#_Toc160010473)

[3.6.1 Time-Series Data 32](#_Toc160010474)

[3.6.2 Random Forest Regression 33](#_Toc160010475)

[3.6.3 Forecasting Evaluation 33](#_Toc160010476)

[3.6.4 Training and Test sets 34](#_Toc160010477)

[3.6.5 Validation and Optimisation 35](#_Toc160010478)

[3.6.6 Forecasting Strategy 36](#_Toc160010479)

[3.7 Ethical Considerations 36](#_Toc160010480)

[4 Results 38](#_Toc160010481)

[4.1 Preliminary Study 38](#_Toc160010482)

[4.2 BERT Classification 38](#_Toc160010483)

[4.2.1 Sample size 38](#_Toc160010484)

[4.2.2 Multi-class vs Binary classification 38](#_Toc160010485)

[4.2.3 BERT Optimisation 40](#_Toc160010486)

[4.2.4 NLP Impact 40](#_Toc160010487)

[4.2.6 Classification Evaluation 41](#_Toc160010488)

[4.2.7 Time-Series generation 42](#_Toc160010489)

[4.3 Feature Relationships and Predictiveness 43](#_Toc160010490)

[4.3.1 Correlational Analysis 43](#_Toc160010491)

[4.3.2 Granger Causality 43](#_Toc160010492)

[4.3.3 Feature sets and Importance Scores 44](#_Toc160010493)

[4.4 Forecasting Performance 45](#_Toc160010494)

[4.4.1 Overall Performance 45](#_Toc160010495)

[4.4.2 Individual Forecasting Performance 48](#_Toc160010496)

[5 Discussion 49](#_Toc160010497)

[5.1 RQ1: Extracting Covid-19 texts 50](#_Toc160010498)

[5.1.1 Class imbalance 50](#_Toc160010499)

[5.1.2 Data Collection 51](#_Toc160010500)

[5.2 RQ2: BERT performance 51](#_Toc160010501)

[5.2.1 BERT suitability 52](#_Toc160010502)

[5.2.2 Factors Impacting Performance 53](#_Toc160010503)

[5.3 RQ3: Correlations and Predictiveness of Reddit Data 53](#_Toc160010504)

[5.3.1 Correlational Insights 53](#_Toc160010505)

[5.3.2 Granger Causality 54](#_Toc160010506)

[5.3.3 Feature Set 54](#_Toc160010507)

[5.4 RQ4: Reddits Impact on Covid-19 Forecasting 55](#_Toc160010508)

[5.4.1 Baseline Comparison 55](#_Toc160010509)

[5.4.2 Forecasting Limitations 56](#_Toc160010510)

[5.5 Future Research 57](#_Toc160010511)

[6 Conclusion 59](#_Toc160010512)

[6.1 Findings 59](#_Toc160010513)

[6.3 Limitations 60](#_Toc160010514)

[6.4 Reflections 61](#_Toc160010515)

[References 62](#_Toc160010516)

[Appendices 67](#_Toc160010517)

[Appendix A 67](#_Toc160010518)

[Appendix B 68](#_Toc160010519)

[Appendix C 70](#_Toc160010520)

[Appendix D 71](#_Toc160010521)

[Appendix E 72](#_Toc160010522)

[Appendix F 73](#_Toc160010523)

# List of Figures

[Figure 1. Overview of Methodological Approach 24](#_Toc159976121)

[Figure 2. Daily frequency of Reddit comments over our timeframe following pre-processing 27](#_Toc159976122)

[Figure 3. Visualising Rolling-Window approach for Training and Test set generation, and variable training size based on forecasting horizon. 35](#_Toc159976123)

[Figure 4. Validation and Training set accuracy for classifying 'Author' dimension labels over increasing sample sizes. 39](#_Toc159976124)

[Figure 5. Mean F1-scores for Multi-class vs Binary classification for 'Contact' dimension labels. 39](#_Toc159976125)

[Figure 6. Mean F1-scores following classification of 'Author' dimensions labels with BERT and RoBERTa classifiers. 41](#_Toc159976126)

[Figure 7. Normalised daily (smoothed) classification counts against normalised U.S.A Covid-19 incidence. 42](#_Toc159976127)

[Figure 8. Validated Feature sets revealing informative features for each respective Training set relative to the forecasting horizon (z). 45](#_Toc159976128)

[Figure 9. RFR model Covid-19 predictions for 7 and 14-day forecasts for all testing periods. 47](#_Toc159976129)

# List of Tables

[Table 1. BERT classification performance for both 'Author' and 'Contact' dimensions on respective holdout test sets. 41](#_Toc159849596)

[Table 2. RFR forecasting performance for the entire testing period for all forecasting horizons. 46](#_Toc159849597)

# **1 Introduction**

## **1.1 Disease Forecasting**

Disease forecasting is crucial for strategic countermeasures and preparedness against infectious diseases [5] and its significance was highlighted by the Covid-19 pandemic, where accurate forecasts informed government policy, public health interventions [1] and ultimately saved lives. Digital epidemiology however has redefined disease forecasting, moving from traditionally structured surveillance/clinical data, and statistical and compartmental models [5] to utilising large-scale real-time digital data with advanced forecasting methods to improve disease forecasting accuracy [4]. This transition to digital data minimises the issue of response lags from acquiring high latency structured traditional data [6] which impacts timely countermeasures to pandemic [3], and its utilisation ultimately helped inform government action to Covid-19 [1].

## **1.2 Digital Disease Forecasting**

Ginsberg’s study [7] was a major milestone for this emerging field, utilising correlating Google searches to forecast Influenza incidence two weeks ahead of CDC reports. This led to Culotta [8] identifying keywords from tweets which correlate and predict CDC reported Influenza incidence. Following this, Google and Twitter became dominant digital resources for disease forecasting [9], with 64% of digital disease surveillance studies utilising Twitter [2].

Despite successfully forecasting various infectious diseases [7, 10-13], Google’s predictiveness remains a function of keyword search volume, and biases like the ‘media effect’ [7] ultimately affect its informativeness without adequate techniques to minimise inherent non-independence [1]. Twitter’s concise texts alongside advances in Natural-Language-Processing however provided a broader methodological landscape, expanding opportunities within digital epidemiology [14, 15], whilst research typically finds Twitter to reveal more valuable insights than Google for disease forecasting [16].

Given this, Twitter has become a significant resource for disease surveillance, providing rapid, real-time indicators into disease prevalence and dynamics, and has been extensively utilised for forecasting Covid-19 [15, 17], Influenza[18-21] and many other infectious disease like Dengue [14], Ebola [22], Zika [23].

Digital resources, notably Twitter, therefore provide a cost-effective alternative to traditional surveillance data [8], particularly significant in geographies with limited screening infrastructure [24].

## **1.3 Reddit for Forecasting**

However, digital epidemiology now faces significant challenges, as Twitter’s recent API paywall restrictions presents a major setback for research, and re-examination of alternative resources is needed for continued resilience against future pandemics. Of these, Reddit emerges as an interesting alternative. Reddit is a discussion focused social media platform, where users exchange advice and information [25] and its in-depth discussions and unrestricted text-limits presents comparative NLP opportunities for epidemiological exploration like that with Twitter [26, 27]. However, its investigation into disease forecasting is limited.

Of these few such studies, Kellner et al. [15] identified that the addition of symptoms from Covid-specific subreddits slightly improves Twitter-enhanced Covid-19 forecasting models, however, it stops short of evaluating Reddit's standalone efficacy in forecasting. Additionally, Kellner’s utilisation of symptoms may not optimally exploit Reddits predictive information, and alternative methodological approaches for information-extraction may be more informative, warranting further investigation.

## **1.4 Aims and Objectives**

This project employs a case study research design, aiming to evaluate Reddits utility as a resource for disease forecasting. We seek to determine if Reddit is a useful alternative to Twitter following recent API restrictions, considering its shared textual nature may reveal comparable predictive indicators to disease, despite its distinct user-base and discourse.  
Our objectives for this research are the following:

* Determine the potential of extracting information from Reddit for disease forecasting, considering its distinct differences to Twitter.
* To review relevant literature and identify the optimal information-extraction methodology for investigating Reddit for disease forecasting.
* To explore and determine relationships between our extracted Reddit information and disease incidence.
* To evaluate the predictive informativeness of our extracted Reddit data in forecasting future disease incidence.

This research contributes to the field of disease surveillance by evaluating Reddit's utility as a forecasting resource, potentially having real-world applications, whilst providing insights into how forecasting methodologies can be applied to Reddit, helping inform and streamline future research.

## **1.5 Research Focus**

To assess Reddits utility in disease forecasting our research focus is limited to Covid-19, due to its global significance generating a wealth of information and widespread screening provides reliable incidence rates for analysis [28]. Additionally, the extensive amount of Covid-19 forecasting research utilising Twitter [15, 29, 30], Google [1, 31], and other internet-based data allows for greater contextualisation of our results, informing our assessment of Reddits utility within disease forecasting.

## **1.6 Report Structure**

Our report is presented in the following structure:

1. Introduction:

Introduces the research field, defining aims and objectives and highlighting our reports significance.

2. Literature Review:

Highlights the research gap of Reddits application within disease forecasting, whilst investigating optimal methodological approaches and methods to investigate Reddit for disease forecasting.

3. Methodology:

Explains our research philosophy, methodological approach and methods used to realise our aim and objectives.

4. Results:

Presents the results of our case study, including the performance of our information extraction approach, correlational and predictive relationships, and our forecasting exercise results.

5. Discussion:

Reviews our findings, with exploration and critiquing of our results, methodologies, and methods relative to each research question, before presenting areas for future research.

6. Conclusions:

Summarising our main findings considering our critique and limitations, then establishing its significance within the broader research field.

# **2 Literature Review**

Research reveals many digital resources to hold extractable predictive information to aid disease surveillance [2], however, limited research has extended this investigation to Reddit despite its similarities.

Reddits text length may be limiting, as shorter texts like Twitters reduces complexity, easier for real-time insights important for disease monitoring [15]. Contrarily, most epidemiological research exploring Reddit is qualitative, utilising Reddits detailed discussions for insights into mental health [26], wellbeing [32], and attitudes to public health interventions [33].

However, outside of epidemiology Reddit is utilised for forecasting, particularly within finance, where subreddit sentiment improves market price predictions [34], highlighting Reddits’ forecasting potential like Twitter’s.

Despite its application outside of epidemiology, Reddits utilisation within disease forecasting is limited. For instance, Sarker [25] aimed to improve the clinical landscape of Covid-19 by extracting symptoms from Reddit using a lexicon-based approach. Building from this, Kellner [15] employed Sarker’s methodology, transforming symptom frequencies into time-series data for forecasting Covid-19, slightly improving Twitter-enhanced models. Kellner’s research highlights Reddits potential for Disease forecasting, but only establishing Reddits’ additive informativeness, not its standalone predictiveness. Additionally, Guo [35] questions Kellner’s validity with respect to the extracted Reddit data, repeating Sarker’s study but ensuring text specificity to Covid-19 using classification, resulting in different extracted symptoms.

Reddits utility within disease forecasting therefore remains questionable, and the appropriate methodology for utilising Reddit for forecasting is not established, which we subsequently investigate.

## **2.1 Information-Extraction**

Keyword frequency analysis is a simple approach for textual information extraction, and Culotta [8] employed this to successfully forecast Influenza using correlating keywords from Twitter. Similarly, Signorini [36] forecasted H1N1 incidence two-weeks ahead using H1N1 keyword frequencies on Twitter. Further research has utilised this approach for many diseases [2], being simplistic and providing real-time analysis, and Kellner’s [15] Reddit symptom-extraction falls under this approach.

Topic modelling, an NLP technique, is comparably utilised for forecasting, where temporal changes in topics/themes correlate and inform forecasting models. Chen generated temporal themes from tweets using probabilistic models to estimate flu trends [37] and Lamsal utilised topic-based latent variables from Tweets to improve Covid-19 forecasts [17]. Topic modelling however is more sophisticated, identifying themes missed through keyword analysis [37], more suitable for Reddits complex text. This is evidenced by Ford's study [38] using Latent-Dirichlet-Allocation (LDA) to capture evolving health-related themes of healthcare workers on Reddit during Covid-19, underscoring the potential extension to capturing Covid-19 themes correlating with Covid-19 incidence on Reddit.

However, both approaches face challenges of non-independence [16] and noise [4] without techniques like classification to improve text specificity before extracting information [8]. Paul [20], following extensive keyword filtering, still employed classification to increase tweet specificity to Influenza before generating influenza-related insights, noting greater validity. Similarly, Guo [35] reinvestigated Sarker's [25] research, using BERT classification to ensure symptom-frequencies are Covid-19 specific, revealing Sarker’s findings as non-Covid-19 specific. However, many keyword and topic frequency/analysis forecasting studies overlook these specificity checks beyond simplistic keyword filtering [39], compromising validity and interestingness. When extracting Reddit data for forecasting, classification is therefore required to increase text specificity.   
  
Sentiment analysis comparably is established within disease forecasting and requires classifiers to categorise text. Das [40] used CNNs to generate and feed Covid-19 Tweet sentiment into regression models for forecasting Covid-19. Similarly, Mohan [41] generated sentiment scores from blogs to improve statistical predictions for Covid-19 in India. Sentiment analysis is investigated often with Reddit [33], and could feasibly be extended to capture correlating sentiments to forecast diseases. However, social media sentiment likewise suffers non-independence, often reflecting sentiment from news coverage [42], questioning its interestingness for disease forecasting beyond simply acquiring news media sentiment.

Interestingly, research has generated predictive insights directly from temporal changes in classified texts indicating an ‘infection’, providing an approximate measure of disease prevalence. Paul [20] revealed classified tweets stating an infection accurately predicted influenza trends, similarly, Shen classified Covid-19 symptomatic users on Weibo, and statistical testing revealed these temporal changes predict Covid-19 incidence 14-days ahead of official reports [24]. Kellner expanded this to additionally classify second-hand and suspecting/ambiguous accounts of Covid-19 infections on Twitter, improving Covid-19 forecasting accuracy against a baseline [15].

This classification-based approach for forecasting provides direct and specific signals to disease [15, 43], eliminating additional information-extraction steps beyond an already required classification task, whilst Guo [35] demonstrated classifying Covid-19 infected ‘individuals’ on Reddit is possible. This classification-based approach is therefore most suitable for our novel investigation of Reddit.

## **2.2 Optimal Classification technique**

For classifying text indicating infections, various techniques have been investigated, from Machine Learning (ML) [8, 39] Deep Learning (DL) [44] and Transformer models [15, 35]. These techniques are critiqued to identify the appropriate classifier for our Reddit text.

Culotta [8] employed simple logistic regression utilising Bag-of-Words representations (BoW) to classify influenza tweets. Though producing a high F1-score (0.9), the BoW feature space required reducing further to limited keywords, simplifying the classification task.Similarly, Santos [39] used SVM classifiers to classify flu-related text, achieving an F-measure of 0.75, but like Culotta [8], following significant feature reduction classification produced a greater F-measure of 0.83. ML classifiers therefore require reduced feature spaces, suitable for shorter less complex classification tasks.

Garzo ́[44] accurately classified Tweets stating infections for several diseases, using vector representations of words with RNN’s with an F1-score of 0.95. This superior performance to [8, 41] suggests semantic and syntactic relationships are important in classifying ‘infected’ tweets, not captured by BoW representation, and ML classifiers lack understanding of context or texts sequential nature which RNN models in [44] captured.

More recently transformer models, which understand context and relationships throughout text, show superior performance in identifying ‘infected’ texts. Kellner [15] used BERT to classify tweets relating to direct and indirect Covid-19 infections, with a weighted F1-score of 0.91. Kellner highlighted BERTs increased complexity and contextual understanding is important for this classification text, noting simpler Naïve Bayes classifiers produced much lower F1-scores.

Although these insights are limited to classifying Tweets, inherently less complex than Reddit text, DL and transformer techniques outperform simpler ML techniques, suggesting classification of ‘infected’ texts benefits from understanding contextual relationships, assumably more significant in Reddits unrestricted length text.

Research has however shown that DL and transformer models can successfully classify disease related Reddit text. Gkotsis [27] used a CNN model to classify Reddit posts along several mental health classifications with high precision and recall for all classes, whilst Guo used BERT to classify Reddit posts to identify ‘authors’ who are Covid-19 positive based on the context and relationships between all posts from each author over time, with high F1 score of 91% [35].

This highlights that DL and BERT can handle the complexity of lengthy Reddit text. BERT however, appears to outperform DL models along various classification tasks in both Twitter and Reddit [28] as it can consider textual context bidirectionally, not just sequentially [45]. Additionally, BERTs ability to utilise text with minimal pre-processing [45], reduces the complexity and steps required to investigate Reddits’ utility in disease forecasting.

## **2.3 Forecasting techniques**

To evaluate our Reddit texts predictive utility, regression modelling is required to forecast Covid-19 incidence [1, 17]. We therefore critique regression techniques within disease forecasting to identify the optimal approach.

Despite advances in ML and DL, traditional forecasting models like ARIMA remain extensively utilised, predicting Covid-19 incidence in India [46], Australia [17], UK [1], USA [15], helping inform public interventions. Although ARIMA struggles capturing non-linear dynamics [17], Saba [47] found ARIMA better predicts Covid-19 incidence than ML models, although limited to periods with linear relationships. Techniques such as Singh’s [48] hybrid ARIMA model however can mitigate ARIMA’s linear requirements, improving Covid-19 prediction errors in Italy by 85%.

However, ARIMA’s linear assumptions and issues with high dimensionality, limits the information and quantity of features for disease forecasting accuracy [49]. ML and DL however are data-driven techniques, effectively processing high dimensional datasets, capturing complex non-linear relationships and assume no prior linearity or assumptions of data distribution. Nor require known transmission dynamics like compartmental models [28], making them ideal for timely predictions of pandemics/epidemics

Regarding ML, whilst early research used simple regression models [8, 20], tree-based approaches and SVMs are common within digital epidemiology [28], regularly outperforming ARIMA. Kane et al. [50] demonstrated Random Forest regression (RFR) outperforms ARIMA for predicting Influenza, with 77% reduction in MSE, Whilst Singh found a Least Square SVM model was 60% better than ARIMA at forecasting Covid-19. Both studies attributed this to ML’s ability to capture non-linear relationships, which better predicted sudden increases in cases, hence ML’s popularity in disease forecasting.

Recent DL techniques have however produced greater forecasting accuracy, better capturing complex temporal dependencies and patterns which ML techniques struggle to capture [28]. For example, Shahid et al. [51] found LSTM effectively captured long-term dependencies, outperforming SVM and ARIMA in predicting Covid-19 in 10 countries. Research such as Keshavamurthy [28] and Kellner [15] further highlights LSTM’s superiority over ARIMA in predicting Covid-19 in Indonesia, and USA respectively. Kellner, utilising Twitter and Reddit, additionally highlights LSTMs long-term forecasting utility, with long-term prediction errors lower than ARIMAs immediate errors, and consistently 70% lower MAE.

However, DL performance within Covid-19 forecasting is inconsistent. Peng predicted Covid-19 incidence globally using Google features, finding RFR consistently outperformed LSTM [31]. Additionally, Ayyoubzadehin [11] forecasted Covid-19 incidence in Iraq with LSTM during the initial pandemic period, finding linear regression outperformed LSTM. Although Ayyoubzadeh highlighted insufficient training data prevented LSTM learning temporal dependencies, even Lamsal [17] found ARIMA outperformed LSTM for forecasting Covid-19 incidence, stating insufficient training data despite containing 618 days.

DL efficacy in forecasting appears dependent on adequate experience to fully utilise these tools. Additionally, DL demands greater computational requirements and requires sufficiently sized data, limiting its immediate use in emerging pandemics [49]. ML techniques like RFR however show greater consistency with easier implementation, and unlike DL’s ‘black box’ approach, RFR provides transparent feature and model insights, useful when understanding disease characteristics and exploring new features [50].

## **2.4 Research Questions**

Our literature review highlights many research gaps in investigating Reddit for disease forecasting but identifies the feasibility of such research. Likewise, we explored and critiqued this research field to identify appropriate methodological approaches/techniques that could be applied to Reddit to realise our research objectives.

This firstly identified classification of ‘infected’ texts as the most suitable information-extraction approach for realising our initial objectives, and determined BERT as superior to ML classifiers [15] for classifying our ‘infected’ Reddit texts. Finally, we establish ML, specifically RFR, as suitable for our forecasting objectives with easier implementation than DL but comparable performance [31, 49].

Our objectives are therefore updated, presented as our Research Questions below which reflect our deductive exploration of the literature:

RQ1. Can Covid-related text be identified from Reddit for a text-classification task?

RQ2. Can BERT accurately classify Reddit texts indicative of Covid-19 infections?

RQ3. Does our classified Reddit text correlate and predict Covid-19 incidence?

RQ4. Does our classified Reddit features inclusion improve Covid-19 forecasting accuracy against a baseline model?

Our research questions ensure a coherent investigation, where each incrementally builds upon and assesses Reddits viability as a resource for disease forecasting, with each addressing a critical aspect of such an effective resource and mirroring related research.

# **3 Methodology**

## **3.1 Research Philosophy**

Unlike a positivist approach to research, relying strictly on evidence, proof, and the scientific method [52], our research embraces post-positivism. Although still requiring measurable data and established methods, our research along with digital epidemiology embraces a pragmatic approach to research [7], strongly acknowledging our limitations, whilst our research’s significance is in its application to improving disease surveillance tools [44] instead of establishing ground truth [52]. A post-positivist philosophy additionally allows us to mix inductive and deductive reasoning, which we do to address our inherently inductive research aim, whilst employing deductive reasoning to select appropriate methodological approaches and methods for investigating Reddit, making our research both exploratory and applied.

This research therefore employs a case study design, ideal for investigatory application of methods and techniques [52], providing flexibility and in-depth exploration of Reddit within disease forecasting. This helps realise our multi-faceted objectives and allows for detailing informative insights and challenges of applying forecasting methodologies to Reddit.

Most research utilising Reddit within Epidemiology is qualitative, where extensive discourse allows for detailed insights [32, 33]. Although informative for revealing factors potentially influencing disease dynamics, forecasting requires quantitative data for empirical evidence-based decision making [1]. Social media-based disease forecasting therefore transforms qualitative data into quantitative representation, and our methodology mirrors this to enable data analysis, and predictive modelling.

Our research takes a retrospective longitudinal design, minimising potentially biased conclusions a cross-sectional focus generates [11]. Our scope is limited to the U.S.A between 1st March to 31st December 2020, capturing the initial Covid-19 pandemic period and aligns with Reddit’s high U.S.A user-base.

## **3.2 Methodology Overview**

Figure 1 provides an overview of our case studies sequential exploration, identifying what stage each research question is addressed.

A diagram of a diagram

Description automatically generated

Figure 1. Overview of Methodological Approach

Figure 1 reveals that following data collection and pre-processing, we label sampled Reddit texts along selected categories relating to Covid-19 infections, addressing RQ1, determining feasibility. These labelled samples then validate/optimise our BERT models before classifying our Reddit dataset, and BERTs performance answers RQ2. Subsequently, classified text is transformed into time-series data, with correlational and predictive analysis addressing RQ3, mirroring related research [5, 8].

Our final stage assesses Reddits predictive utility, evaluating Reddit-enhanced RFR Covid-19 predictions against baseline models following RFR validation/optimisation. For this forecasting exercise we employ a Rolling-Window approach for generating training and test sets, a Walk-Forward testing approach, and lagged Covid-19 features to forecast future incidence. This produces an extended testing period, regular model updates, mimicking real-world forecasting scenarios in-line with comparable disease forecasting research [15, 31, 53].

## **3.3 Data Sources**

### **3.3.1 Covid-19**

Covid-19 data was acquired from WHO's public dataset, available at https://data.who.int/dashboards/covid19/data?n=c., containing globally compiled data via aggregated laboratory-confirmed counts reported to WHO headquarters [54]

The dataset was filtered to select only new confirmed Covid-19 cases in the U.S.A between our investigative timeframe.

### **3.3.2 Reddit Data**

Previous research [15, 25, 35] utilised PushShift.API for Reddit data collection, however recent Reddit paywall restrictions have blocked this tool. Instead, an opportunistic pre-collected Dataset ‘Reddit COVID Dataset’ was acquired from ‘SocialGrep’ (www.socialgrep.com), consisting of 17.7 million Reddit comments containing the word ‘COVID’ prior to October 25, 2021, aligning with our research aim [55]. The dataset is anonymised and full compliance with License CC BY 4.0 was followed.

### **3.3.3 Pre-Processing**

The dataset was filtered to select relevant features - text, timestamp, subreddit name, unique ID - and filtered to our study's temporal scope. Reddit text was cleaned, removing excessive whitespace, newline characters, non-alphanumeric symbols, and URLs.

Reddit lacks geospatial data, however subreddit names reflect specific themes. We therefore increased specificity to the U.S.A by filtering subreddits with keywords indicative of American locations, i.e. States, Cities, Counties. Additionally, following preliminary analysis, Covid-19 specific subreddits were included to improve target class representation. Non-USA Covid-specific subreddits, i.e. 'r/CoronavirusCanada', were removed whilst subreddits with no-geospatial indication were retained due to high representation of our target text and some specificity to the U.S.A [15]. Keyword-filtering retrieved 673 subreddits from an initial 75,465.

Although ML classifiers require NLP pre-processed text [39], this may remove informative content for BERT [45]. Additionally, BERTs respective ‘pre-processor’ effectively handles text processing, therefore NLP pre-processing was not applied [45].

To avoid truncating texts exceeding BERT’s 512-token limit, we split texts exceeding 1280-characters (2.5 character/token estimate) into independent chunks [35]. Larger BERT variants were considered but have significant computational demands. Finally, split-texts were tokenized and texts still exceeding the 512-token limit (0.16%) were excluded, leaving 337,128 instances, with daily frequency visualised in Figure 2.

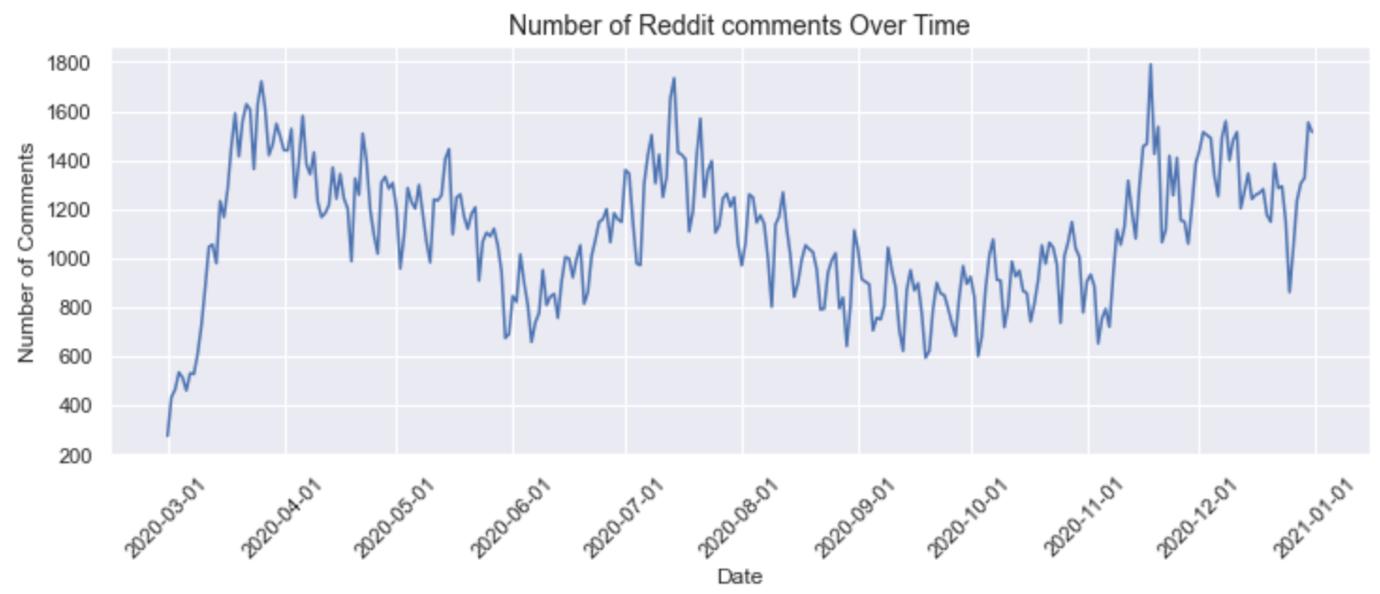


Figure 2. Daily frequency of Reddit comments over our timeframe following pre-processing

## **3.4 Reddit Information Extraction**

Classifying text indicative of a personal infection is informative for disease forecasting [20, 24]. Kellner however additionally captured text stating second-hand and suspected/ambiguous Covid-19 infections [15]. This approach captures a greater amount of Covid-infection related text of direct and indirect Covid-19 transmission pathways, whilst suspected/ambiguous texts identify underreported cases potentially preceding confirmed infections as authors speculate prior to testing.

Our Reddit text is therefore classified along an ‘Author’ and ‘Contact’ dimension.

With the ‘Author’ dimension containing labels ‘YES’, ‘MAYBE’, and the ‘Contact dimension containing labels ‘Work’, ‘Family’ and ‘Acquaintance’, based on Kellner’s research [15].

Label: **YES**Definition: Indicating/implying a Covid-19 infection.Examples:

*“…I tested positive for Covid…”*

“*…tested positive last week too…”*

*“…thought it was allergies. It was covid.”*

Label: **MAYBE**Definition: suggesting/suspecting or ambiguous to having a Covid-19  
 infection.Examples:

*“…Odds I have covid? I’m thinking 90%…”*

*“…my work they assume I have covid…”*

*“…perhaps mild symptoms of COVID, but didn't get tested…”*

Label: **FAMILY/ WORK/ACQUAINTANCE**Definition: Indicating/implying a Covid-19 infection of a family member/colleague/acquaintance  
Examples:

*(‘FAMILY’) “My mom has covid now…”*

*(‘WORK’) “…I contracted this from my place of employment…”*

*(‘ACQUAINTENCE’) “…someone I live with has been covid positive…”*

The ‘Author’ and ‘Contact’ dimensions are not mutually exclusive, Reddit text is therefore labelled along each dimension, and a TKinter Python GUI developed to aid classification.

## **3.5 Text Classification**

### **3.5.1 BERT**

BERT (Bidirectional Encoder Representations from Transformers) is a transformer-based DL model [45], selected for its capability to understand textual context and relationships. Vasmani et al. [56] introduced the transformer architecture, innovating attention mechanisms to focus attention over various texts simultaneously. BERT modifies the transformers encoder to processes text bidirectionally, producing greater understanding of complexity and nuances in language [45].

Additionally, BERTs extensive pre-training makes it versatile for various NLP tasks, including classification, providing fine-tuning capabilities via supplementing task-specific data for specialised NLP tasks [17]. BERT therefore minimises training requirements respective to conventional models, useful during pandemics where urgent deployment is essential [15].

We used ‘bert\_en\_uncased’ model from TensorFlow Hub, 4th version, with 12 layers, 768 hidden-layers, and 12 attention heads, and respective ‘bert\_en\_uncased\_preprocess’ pre-processor.

We utilised Google Colab as our computational environment to utilise T4 GPUs for BERT training.

### **3.5.2 Classification Evaluation**

For classificationevaluationthe following metrics are employed:

Precision:

* Measures the proportion of correctly classified positive texts from all positive texts, defined below:

Recall:

* Measures the proportion of actual positive texts correctly classified by the model, defined below:

F1-score:

* Balances Precision and Recall, harmonising as a single metric, useful for imbalanced datasets, and defined below:

For multi-class classification we average Precision, Recall, and F1-Score over all classes, employing the weighted average accounting for class imbalance.

### **3.5.3 BERT Validation and Testing**

Our initial BERT model employs Adam optimizer (learning rate=2e-5), batch size=16, a classification-head containing a dropout (rate=0.2) and dense layers tailored to the classification task (binary/multi-class) [57].

We implement stratified 10-fold cross-validation, improving reliability and generalisation, whilst balancing class representation equally. Class weights are calculated inversely proportional to frequency for each iteration, epochs are set to 10 allowing gradual iterative learning, and early stopping based on validation performance reduces overfitting [57].

Classification performance over increasing labelled sample sizes with the aid of learning curves determined the optimal sample size. Additionally, a multi-class versus binary classification approach is evaluated for classifying the ‘Author’ and ‘Contact’ labels, determining the optimal approach.

Hyperparameters were fine-tuned sequentially. Early stopping metric was evaluated between validation cross-entropy and accuracy, followed by learning rates (1e-5, 2 e-5, 5e-5) and dropout rates (0.2-0.5). Then the neural network architectures were optimised, investigating the performance of additional dense and batch normalization layers.

After tuning/optimising, each model was validated against a 20% stratified holdout test set to assess generalisation before classifying our Reddit dataset.

To contextualize BERTs performance and validate our assumptions we compared our model against RoBERTa, a more sophisticated BERT-variant [58], and against a BERT model trained on our NLP stemmed and lemmatised sample.

## **3.6 Forecasting**

### **3.6.1 Time-Series Data**

Our Reddit dataset is classified by both BERT models. The frequency of each classification label is aggregated by day and merged with U.S.A Covid-19 incidence, generating our time-series dataset.

To reduce label overrepresentation from our text-splitting approach, duplicate instances based on ‘id’ and classification label were removed. We then engineered features from each classification label to capture more information, including rolling means, standard deviations, and differencing [53].

Additionally, we lag the Covid-19 incidence by 1-day, generating a “Traditional” autoregressive feature whilst allowing for generating comparative autoregressive baseline models, consistent with typical disease forecasting research [1, 15, 30].

All features for forecasting are normalised.

Finally, for our forecasting objectives, most comparable research investigates disease incidence from 1-14 days ahead [1, 15], representing the typical latency delay of clinical/laboratory confirmed incidence [3]. We therefore create lagged features, shifting the Covid-19 incidence back 7 and 14-days, serving as respective target variable for our 7 and 14-day forecasting models, allowing RFR to identify relationships between current-day features and future Covid incidence [53].

A description of our time-series features is in Appendix-A, Figure 1.

Spearman Rank correlation identified immediate and lagged relationships between Reddit features and Covid-19 incidence over 28-days, whilst cross-correlation investigated feature independence to mitigate multicollinearity which can impact model training [31]

For preliminary predictive assessment of our classification features to Covid-19 incidence, we employed Granger Causality tests over a 28-lag, following differencing of our features to meet stationary requirements.

### **3.6.2 Random Forest Regression**

RFR [59] is an ensemble tree-based method, constructing many individual decision trees whose regression value is the mean of all outputs. This ensemble approach makes RFR robust in predictive modelling and preventing overfitting, with greater generalisation than individual decision trees [49].

RFR is resilient to high-dimensionality due to efficient feature selection, identifying and focusing attention to features with the highest information gain. This capability reduces overfitting to irrelevant noisy features whilst providing feature importance scores which quantifies each features predictive utility to the model [59].

Disease forecasting involves complex variable interactions/relationships and non-linear dynamics, but decision trees can fit non-linear patterns. Since RFR’s output is the aggregation of many decision trees, RFR can model these complex non-linear interactions, therefore suitable for disease forecasting [59].

### **3.6.3 Forecasting Evaluation**

RFR is evaluated with the error metrics below, and for formulas; *Ai and Pi* are actual and predicted values, *n* is total observations.

Mean Absolute Error (MAE):

* Reports average error over all predictions against actual values, with errors weighed equally, defined below:

Root Mean Squared Error (RMSE):

* Weighs larger errors greater, highlighting the variance, squaring errors before averaging and applying the root, defined below:

Mean Absolute Percentage Error (MAPE):

* Provides a scale-independent metric, normalising errors to the actual values scale, allowing comparisons between periods with different target variable scale. Formula is defined below:

### **3.6.4 Training and Test sets**

We employed a Rolling-Window approach for generating Training and Test sets, with Test sets chronologically following Training sets, aligning with time-series forecasting methodologies [15], visualised in Figure 3.

This created 9 consecutive Test sets each containing 28-days between 23rd April - 31st December 2020. Each Test set has three respective training sets preceding it, representing the same training sets but with the immediate 1, 7, or 14 days preceding the respective test set excluded relative to the forecasting horizon ‘*z*’ (1, 7, and 14 days) each training set represents, preventing data leakage and simulating a real-world forecasting scenarios [53], generating 27 training-sets (T1-27).

A screenshot of a test

Description automatically generated

Figure 3. Visualising Rolling-Window approach for Training and Test set generation, and variable training size based on forecasting horizon.

### **3.6.5 Validation and Optimisation**

Each training sets features are validated to improve RFR generalisation [53].

‘TimeSeriesSplit’ from sklearn was employed for time-series cross-validation, maintaining temporal integrity. Our validation procedure generated average importance scores over 10-fold cross-validation, followed by ablation studies where top *n* features are inputs for training over 10-fold validation, identifying optimal feature sets based on minimising MAE and RMSE, mirroring Peng’s feature-validation [31].

RFR hyperparameters are optimised for each training set using GridSearch and 10-fold cross-validation, with parameters; ‘n\_estimators’ (100-300), ‘max\_depth’ (10-30), ‘min\_samples\_leaf’ (2,5,10), ‘min\_samples\_split’ (1,2,4), ‘criterion’ (MSE, MAE).

### **3.6.6 Forecasting Strategy**

Since we evaluate Reddits informativeness against baseline models, training sets contain at least one Reddit feature following feature validation, regardless of ablation studies revealing no features improve MAE or RMSE. Additionally, for baseline comparison each training set period is replicated, but the feature set contains only the autoregressive “Traditional” previous Covid-19 incidence, with RFR hyperparameters again optimised.

For testing, a Walk-Forward approach was utilised, where models predict the following test instance, before addition to the training set for model-retraining, and repeated, consistent with time-series forecasting methodologies, replicating real-world forecasting scenarios with regular model-updates with new information [53].

## **3.7 Ethical Considerations**

Though Reddit users post content publicly, issues of informed consent arise since users are unaware their posts are utilised for subsequent research and potential real-world applications. This likewise links to privacy issues, as textual exploration may inadvertently identify individuals who have a right to anonymity [2].

Our research therefore maintains the minimum amount of potentially identifiable information, and we present our results only following data aggregation, reducing privacy risks [60].

Whilst our research and the disease surveillance field aims to benefit public health with principles such as beneficence and non-maleficence [2], potentially inaccurate and biased methodologies, along with underrepresented users and communities within our Reddit data, may lead to skewed and potentially harmful forecasting conclusions [60]. To minimise harm, our discussion will critically dissect our methodology and findings, addressing the limitations of our forecasting research, minimising misleading harmful conclusions.

Considering these ethical issues, fast-track ethical approval was obtained from the University of York’s Departmental Ethics Officers, and adherence to ethical best practice was followed.

# **4 Results**

## **4.1 Preliminary Study**

Firstly, our preliminary study of 500 labelled samples following USA-specific subreddit filtering identified large classification imbalance. Subsequent addition of COVID-specific subreddits improved class imbalance for all labels; Yes=14%, Maybe=11.1%, Family=7.46%, Acquaintance=3.54%, Work=1.06%, leading our decision to include Covid-specific subreddits in our dataset.

## **4.2 BERT Classification**

### **4.2.1 Sample size**

Figure 4. visualises the learning curve plotting training and validation accuracy. Initial tests show overfitting, with high training accuracy and lower validation accuracy of 0.76. Increasing sample sizes reduced overfitting, but minimal increase in validation accuracy was observed from 2574 to 4291 samples (0.79 and 0.8 respectively), our final sample size therefore remained at 4291 due to diminishing returns.

### **4.2.2 Multi-class vs Binary classification**

Evaluation between multi-class versus segmented binary classification for our ‘Author’ ‘YES’ and ‘MAYBE’ labels revealed no improvement in classification performance, shown in Appendix-B, Table 1.

Whilst for our ‘Contact’ labels, a Binary classification approach, where all labels are re-labelled a single ‘CONTACT’ label, revealed superior performance as shown in Figure 5, with an F1-score of 0.76. Moving-forward, our ‘Contact’ dimension, is a binary classification problem, addressing the class imbalance affecting multi-class performance.

A graph of a graph showing the size of a number of scores

Description automatically generated with medium confidence

Figure 4. Validation and Training set accuracy for classifying 'Author' dimension labels over increasing sample sizes.

A graph of different colored bars

Description automatically generated with medium confidence

Figure 5. Mean F1-scores for Multi-class vs Binary classification for 'Contact' dimension labels.

### **4.2.3 BERT Optimisation**

Hyperparameter and architectural tuning was optimised as stated in the methodology, with plots of Hyperparameter and Architectural tuning presented in Appendix-B, Figures 1-4.

* Early stopping metric 'validation accuracy' produced higher F1-scores with less variation for both ‘Author’ and ‘Contact’ dimensions.
* Learning rate 2e-5 and dropout rate 0.2 produced the highest F1-scores and greater classification stability for both BERT models.
* Default classification-head (Arch-1) outperformed additional dense layers (Arch-2) and Batch Normalisation layers (Arch-3) for all performance metrics for both BERT models.

### **4.2.4 NLP Impact**

NLP pre-processing (stemming and lemmatization) decreased BERT classification performance for the 'Author' dimension, with F1-scores for 'YES', 'MAYBE', and 'NO' labels decreasing by 0.04, 0.02, and 0.04, respectively, and impairing classification stability, supporting our assumption against NLP pre-processing. Results are visualised in Appendix-2, Figure 5.  **4.2.5 RoBERTa Comparison**

RoBERTa outperformed BERT for classifying our ‘Author’ labels over all metrics, improving F1-scores by 0.07 and 0.08 for the ‘YES’ and ‘MAYBE’ labels, as shown in Figure 6. However, RoBERTa required 572 minutes for classification, 300% longer than BERT. Due to time and resource constraints BERT was preferable.

**A graph of different colored rectangular shapes

Description automatically generated**

Figure 6. Mean F1-scores following classification of 'Author' dimensions labels with BERT and RoBERTa classifiers.

### **4.2.6 Classification Evaluation**

BERT showed respectable performance in classifying personal Covid-19 infections ('YES'), achieving a reasonably high and balanced precision (0.76), recall (0.72), and F1-score (0.74). For suspected/ambiguous references to an infection (‘MAYBE’), the model performed poorer, notably precision, achieving a precision of 0.58, recall of 0.75, and F1-score of 0.65, indicating more noise is present in the ‘MAYBE’ time-series data.

BERT also showed respectable performance identifying second-hand accounts of infections (‘CONTACT’) with reasonable precision (0.7), high recall (0.83) and F1-score of 0.75. Though a higher F1-score than ‘YES’ text, ‘CONTACTs’ lower precision introduces more noise into our time-series data.

Both BERT models performed well in identifying text unrelated to accounts of infection, as shown in Table 1.

### **4.2.7 Time-Series generation**

Following BERT classifying our Reddit dataset and removing duplicates due to our text splitting procedure, the following classification counts are:

‘YES'=23,799, 'MAYBE'=25,140, 'CONTACT'=26,809.

Time-series data was generated as defined in the methodology, with daily counts visualised in Figure 7.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Label** | **Precision** | **Recall** | **F1-Score** |
|  |  |  |  |  |
| **‘Author’** | **YES** | 0.76 | 0.72 | 0.74 |
|  | **MAYBE** | 0.58 | 0.75 | 0.65 |
|  | **No** | 0.92 | 0.84 | 0.88 |
| **‘Contact’** | **CONTACT** | 0.7 | 0.83 | 0.75 |
|  | **No** | 0.97 | 0.94 | 0.96 |

Table 1. BERT classification performance for 'Author' and 'Contact' dimensions on holdout test sets.

A graph showing the number of the number of individuals

Description automatically generated with medium confidence

Figure 7. Normalised daily (smoothed) classification counts against normalised U.S.A Covid-19 incidence.

## **4.3 Feature Relationships and Predictiveness**

### **4.3.1 Correlational Analysis**

Figure 7 reveals relationships between our Reddit labels and Covid-incidence. Spearman rank identifies a strong correlation of 0.95 for 'YES', a reasonably strong correlation of 0.70 for 'CONTACT', whilst a weak correlation of 0.37 for 'MAYBE'. Therefore, personal mentions of Covid-19 infections strongly align with Covid-19 incidence, and sharp increases in ‘MAYBE’/‘CONTACT’ frequency during the initial pandemic period seen in Figure 7 likely weaken correlations.

Lag correlations with Spearman rank revealed highest correlations at lag=0, indicating no time-delayed relationships for all labels, suggesting stronger immediate relationships to the Covid-incidence. The ‘YES’ labels strong correlation over the 28-day lag however suggests potentially inconsistent lagged relationships. In contrast, 'MAYBE' has consistently weak/declining relationships over time, shown in Appendix-C, Figure 1.

Spearman cross-correlation revealed high correlations between smoothed ‘CONTACT’ and ‘MAYBE’ labels (0.83), but lower relationships to the ‘YES’ label, 0.64 and 0.32 respectively. Only smoothed label features were retained moving-forward, to reduce multicollinearity and this was later validated by our RFR validation process.

### **4.3.2 Granger Causality**

Granger causality over 28-day lag revealed predictive utility of our features, visualised in Appendix-C, Figure 2.

’YES’ text has significant immediate and time-delayed predictiveness at lags 1-2 and from lag 25 (p<=0.05), with inconsistent short-term predictiveness. ‘CONTACT' and 'MAYBE' demonstrate comparable strong predictiveness between lags 9-22, indicating consistent short/medium-term time-delayed informativeness.

### **4.3.3 Feature sets and Importance Scores**

Validated ‘Reddit+Traditional’ feature sets are visualised in Figure 8, and Appendix-D, Figure 1, presents average features importance scores.

At immediate forecasting (z=1), previous-days Covid-19 incidence is consistently most informative (Importance=0.52), with ‘YES’ showing some importance (0.33) whilst other Reddit features add little information to the autocorrelating feature alone.

At increasing horizons (z=7, 14), Reddit labels become increasingly informative for predicting Covid-incidence. At z=14 ‘YES’s importance increases to 0.41, greater than previous Covid-incidence of 0.22, and often most informative.

Additionally, ‘MAYBEs’ importance increases with forecasting horizon (0.18), notably more important than ‘YES’ at initial pandemic-periods whilst increasingly present within feature sets. ‘CONTACT’ however is never within the top two most informative features and represented less in feature sets.

Engineered features show minimal importance, with highest importance being 0.03 at z=14.

A chart with different colored squares

Description automatically generated

Figure 8. Validated Feature sets revealing informative features for each respective Training set relative to the forecasting horizon (z).

## **4.4 Forecasting Performance**

‘T-model’ (baseline) is trained with only ‘Traditional’ (previous-days Covid-incidence) features. ‘RT-model’ (Reddit-enhanced) is trained with ‘Traditional’ plus respective Reddit feature set from Figure 8, and *z* represents the forecasting horizon (1,7,14-days). Figure 9 visualises the Covid-19 predictions for both models at 7 and 14-day forecasts.

### **4.4.1 Overall Performance**

Table 2 presents our ‘RT’ and ‘T-model’ overall forecasting performance.

For immediate forecasting, ‘T-model’ outperforms the ‘RT-model’ (MAE=11,380 vs 15,303, RMSE=17,171 vs 22,629 respectively), implying Reddit features reduce RFRs overall generalisation at immediate forecasts.

However, with increasing forecasting horizons, ‘RT-models’ demonstrate notable improvements respective to ‘T-models’. At z=7, ‘RT-models’ outperforms ‘T-models’ with lower MAE (31,118 vs 40,551) and RMSE (46,939 vs 56,488), a relative error reduction of 23.26% and 16.9% respectively.

This trend continues at z=14 (MAE=43,050 vs 63,150, RMSE=82,337 vs 104,387), with the 'RT-model' showing a relative error reduction of 31.83% and 21.12% for MAE and RMSE to the 'T-model', indicating Reddit features become increasing informative at larger forecasting horizons, and highlighting autocorrelation's diminishing forecasting informativeness at longer horizons.

With increasing forecasting horizons, RMSE for both models increases greater than MAE, suggesting more extreme prediction errors, although less extreme in 'RT models’.

MAPE values support the MAE and RMSE results, revealing 'RT-models' outperform ‘T-models’ at increasing forecasting horizons, producing lower MAPEs at z=7, 6.61%, and z=14, 8.40%, against 'T-models' 10.53% and 15.15%, respectively.

Interestingly, at z=14, the RT-model's MAE is comparable to the 'T-models' MAE at z=7, only increasing by 6.16%. Without Reddit features, the baseline ‘RT-models’ MAE increases by 55.7% from z=7 to z=14. This is similarly observed with the MAPE metric, with the RT-model’ having lower MAPE at longer forecasting horizons (z=14) than baseline models at shorter horizons (z=7), 8.40% and 10.53% respectively, further highlighting the validated Reddit features reveal time-delayed predictive information to the Covid-19 incidence.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **RFR** | **1-Day ahead** | | | **7-Day ahead** | | | **14-Day ahead** | | | |
|  | **MAE** | **RMSE** | **MAPE** | **MAE** | **RMSE** | **MAPE** | **MAE** | **RMSE** | **MAPE** | |
| **‘T’** | 11,830 | 17,171 | 2.6% | 40,551 | 56,488 | 10.5% | 63,150 | 104,387 | | 15.2% |
| **‘RT’** | 15,303 | 22,629 | 2.9% | 31,118 | 46,939 | 6.6% | 43,050 | 82,337 | | 8.4% |

Table 2. RFR forecasting performance for the entire testing period for all forecasting horizons.

A graph of different colored lines

Description automatically generated

Figure 9. RFR model Covid-19 predictions for 7 and 14-day forecasts for all testing periods.

### **4.4.2 Individual Forecasting Performance**

As evident in Figure 9, forecasting performance is inconsistent between test sets. Appendix E, Table 1, presents individual RFR performance to each test set for each forecasting horizon.

At z=1, 'RT-models' outperform ‘T-models’ at Test sets 1,2,4,5,6, representing the earlier pandemic period, however, ‘T-models’ outperform ‘RT-models’ at test sets 7,8,9, representing periods of sharp increases in Covid-incidence, notably at Test set 9 with MAE and RMSE 67.26% and 65.13% lower respectively. Overall, at z=1, Reddit features add limited information restricted to stable pandemic periods.

At z=7, 'RT-models' outperform 'T-models' in Test sets 1-7, again representing earlier pandemic periods and now both initial Covid 'wave' periods (Test sets 3,7), evident in Figure 10, with notably over 50% lower MAE and RMSE for Test sets 2,4,5.

Additionally, 'RT-models' show greater consistency in prediction, with lower MAPE standard deviation of 3.19% against ‘T-models 5.23%, and 'T-models' showing more frequent test sets with MAPEs exceeding 10%. However, 'T-models' again perform better in Test sets 8 and 9, revealing Reddit features provide limited information during rapidly increasing Covid-incidence.

By z=14, 'RT-models' generally outperform 'T-models', including Test sets 3 and 7 representing initial Covid ‘waves’, and are more consistent with lower MAPE standard deviation of 4.74% against 'T-models' 10.18%.

'RT-models' now outperform 'T-models' in test sets (8 and 9) at larger forecasting horizons, based on RMSE and MAPE, however, any improvement is minimal, and reaffirms the little information Reddit features provide at this rapidly evolving pandemic period.

# **5 Discussion**

Our exploration revealed Reddit as an informative resource for Covid-19 forecasting. We demonstrate our information-extraction approach of identifying personal, suspected/ambiguous, and second-hand accounts of infections hold time-delayed predictive utility for Covid-19 forecasting, though personal accounts are most consistently strong predictors. Correlational analysis and Granger testing revealed predictive relationships, whilst our forecasting exercise validated Reddits increasing predictive value over increasing forecast horizons. Our forecasting exercise however revealed inconsistent predictiveness, showing notable informativeness during initial Covid waves and stable pandemic periods, but limited value during the pandemic peak from 5th November to 31st December.

Our novel application of disease forecasting methodologies to Reddit revealed unique challenges, and we provide pragmatic solutions addressing Reddits’ lack of geospatial data and unrestricted text length. Our research highlights BERTs capability with classifying Covid-specific Reddit text and RFRs utility in forecasting Covid-19 incidence whilst providing transparent insights to our investigative Reddit features.

We dissect our results and methodological approaches relative to each research question below, then address avenues for future research.

For simplicity/clarity, ‘YES’/‘MAYBE’/‘CONTACT’, when referring to literature corresponds to equivalent classification categories to ours.

## **5.1 RQ1: Extracting Covid-19 texts**

Like established digital platforms, we revealed Reddit contains sufficient text indicative to COVID-19 infections for classification and forecasting, following increasing class representation by including Covid-specific subreddits. Sarker and Guo [25, 35] inspired this decision, revealing Covid-specific subreddits contain frequent texts indicating infections, whilst maintaining some USA geographical specificity as Kellner successfully forecasted USA Covid-19 incidence with extracted Covid-subreddit symptoms in combination with Twitter data without accounting for geographical specificity [15].

Class representation was ‘YES’:14%, MAYBE’:11.1% and CONTACT’:12% (after re-labelling to a single ‘Contact’ label to enhance BERT performance). This representation is consistent with related research, highlighting Reddits’ feasibility for information extraction comparable to established platforms.

For Covid-19, Kellner found a representation of ‘YES’:14.8% on Twitter [15] and Shen a representation of 5.7% on Weibo [24], whilst Garzo ́n-Alfonso [44] identified a representation of 63.3% for Influenza on Twitter.

Our ‘MAYBE’ and ‘CONTACT’ labels have fewer comparative literature, but Kellner [15] found comparable representation of 9.6% and 23.4% respectively, albeit a higher representation of ‘CONTACT’ text.

### **5.1.1 Class imbalance**

To improve class representation, we could have alternatively applied Garzo and Santos’s approach [39, 44], both generating high class representations of Influenza ‘YES’ texts on twitter of 63.3% and 35.1% respectively, by extensively filtering text with disease-specific keywords. This returns fewer but more specific texts related to infections, however, our approach aimed to maximise the amount of available data for predictive insights. Santos [39] however demonstrates a smaller highly-representative dataset of 2704 tweets is sufficient for Influenza forecasting, suggesting we could have employed keyword-based text filtering instead to increase class representation, maintaining specificity to the USA, whilst a smaller dataset would be sufficient for predictive insight.

### **5.1.2 Data Collection**

Additionally, our pre-collected dataset contains texts matching a single keyword 'COVID', however comparable research typically employs larger sets of disease-specific keywords for data collection [8, 24, 29]. Our dataset is therefore a subset of our intended dataset, likely excluding relevant text for this research. Limitations regarding class representation or predictive insight may therefore relate to insufficient data collection rather than inherent limitations to Reddit.

## **5.2 RQ2: BERT performance**

Our BERT models classified Reddit text reasonably effectively, producing F1-scores of 0.74, 0.65 and 0.75 for ‘YES’, ‘MAYBE’ and ‘CONTACT’ labels respectively.

Our 'YES' label, indicating personal infections, represents our most direct feature to disease incidence, and BERTs effectively robust at identifying it with balanced precision and recall and the highest precision. Additionally, for second-hand accounts of infection BERT performed well, achieving the highest recall (0.83) indicating BERTS capability to distinguish text stating/inferring an infection regardless of being a personal or second-hand account.

BERT however showed poorer accuracy for ambiguous/speculative mentions of Covid-19 infections. Its poorer precision of 0.58 indicates BERT struggled to distinguish ambiguity, introducing considerable noise into our time-series dataset.

### **5.2.1 BERT suitability**

BERTs performance slightly underperforms against research. Kellner’s [15] BERT model produced F1-scores of 0.85 and 0.82 for ‘YES’ and ‘CONTACT’ classes, but poorer ‘MAYBE’ performance of 0.62 also highlighting difficultly in classifying ambiguous text. However, simpler ML techniques outperform our BERT models, with Santos [39] classifying influenza ‘YES’ tweets with an F-measure of 0.83 using SVM and Naïve Bayes, and Shen [24] classifying Covid-19 ‘YES’ Weibo posts with F1-scores of 0.88 for SVM and Random Forest classifiers.

Comparisons to related research however is not directly comparable, as our larger Reddit text (mean=523 characters) makes classification more complex [35], requiring more complex models for equivalent performance. This is supported by our more-complex RoBERTa models comparison, showing superior performance to BERT across all labels, more in-line with comparable research. However, Shen’s [24] ML classifiers outperformed Kellner’s [15] more-advanced BERT model in classifying shorter Covid-19 text. The effectiveness of simpler classification techniques indicates the classification task for shorter text is potentially more straightforward, and BERT’s complexity may impair performance.

### **5.2.2 Factors Impacting Performance**

We labelled 4,291 texts, after learning curves revealed minimal generalisation gains with larger samples, whilst Kellner [15] and Shen [24] labelled significantly more, 6,027 and 11,575 texts respectively. Assuming their sample size was based on optimal classification generalisation and shorter text being less complex suggests our BERT model requires more labelled samples for training to generalise accurately.

However, unlike comparable research utilising multiple annotators and majority/consensus techniques, we had a single annotator, therefore higher risk of bias and susceptibility to inconsistent label definitions [24] particularly with our ‘MAYBE’ label which captures ambiguous text. Multiple annotators/consensus techniques may improve classification performance, producing more valid data for subsequent forecasting.

## **5.3 RQ3: Correlations and Predictiveness of Reddit Data**

### **5.3.1 Correlational Insights**

Correlational analysis revealed strong associations between text stating personal infections and Covid-19 incidence (0.95), comparable to related literature, with Kellner [15] showing a correlation of 0.94 and Didi [30] a weaker correlation of 0.64 for Covid-19 ‘YES’ tweets and Covid-incidence. Also consistent with correlations identified in other researched diseases, 0.89 for Influenza ‘YES’ tweets and 0.94 for H1N1 (‘YES’) tweets [61, 62]. Reddit therefore reveals direct signals related to Covid-19 prevalence, comparable toTwitter**,** notable considering our lower classification precision.

Our 'CONTACT' features lower correlation (0.70) is comparable to referenced literature, but 'MAYBEs' correlation of 0.37 is significantly lower, noticeably lower than Kellner's (0.85)[15]. Though comparable F1-scores, our ‘MAYBE’ labels lower precision than Kellner’s (0.58 vs 0.79 respectively), resultingly introduces more false positives, impacting ‘MAYBEs’ true correlation to Covid-19 incidence.

### **5.3.2 Granger Causality**

In contrast to lag-correlations, granger tests revealed ‘MAYBE’ and ‘CONTACT’ as significant predictors at increasing lags, suggesting time-delayed non-linear relationships to Covid-19 incidence. Interestingly, the ‘YES’ feature showed inconsistent predictiveness. Kellner [15] however identified Granger causality from lag-1 for all features at lower significance values (0.001), suggesting our Reddit features have less predictive power than comparable Tweets.

### **5.3.3 Feature Set**

Validated feature sets reveal 'YES' as consistently and ‘MAYBE’ increasingly important to RFR models for predicting future Covid-incidence, supporting Granger causality tests. However, unlike related studies [15, 30], our Reddit features add little nowcasting information.

Despite 'MAYBEs’ notably lower precision/correlation it’s increasingly revealed as an informative predictor, either enough true information remains represented, or noise has introduced confounding variables.

## **5.4 RQ4: Reddits Impact on Covid-19 Forecasting**

Our forecasting exercise reveals Reddit features provide relative MAE improvements of 23.26% at z=7, and 31.83% at z=14 against baseline models.

Though slightly lower, this aligns with comparable literature forecasting Covid-19, with Kellner [15] demonstrating a 39.57% (z=7) and 30.1% (z=14%) reduction in MAE from baseline LSTM models with comparable Twitter classification features, and Lampos [1] a greater MAE reduction of 29.5% (z=7) and 43.61% (z=14) using Google volumes with ARIMA(X). Our relative model improvement is also consistent with research investigating other diseases, with Paul [20] demonstrating Influenza ‘YES’ tweets improve baseline autoregressive models by 27.5% (z=7) and 24.5% (z=14). Overall suggesting our Reddit features are similarly effective for Covid-19 forecasting, as Twitter and Google features.

At granular inspections however, informativeness is poorer at a significant pandemic peak (test sets 8-9) questioning Reddits reliability for Covid-19 forecasting,

Few other studies like Chatterjee's [63], who investigated Covid-19 forecasting in India dissect their forecasting performance at specific periods, most summarise performance over entire testing periods [1, 15, 30], potentially masking periodic poorer performance. It’s therefore difficult to compare our models’ reliability during specific periods, though doing so is important as accurate modelling during pandemic ‘waves’ is critical for timely interventions [3].

### **5.4.1 Baseline Comparison**

Our forecasting exercise frames Reddits’ predictive-informativeness relative to baseline ‘autoregressive’ models, trained on historical Covid-incidence. Although some studies evaluate performance relative to models trained with other established resources like Google and Twitter [16, 39], contextualising each resources informativeness, our approach is consistent with disease forecasting research [1, 15, 30] producing a standardised relative measure of improvement, providing comparisons between research.

### **5.4.2 Forecasting Limitations**

Despite RFRs robustness, employing a single forecasting technique is inconsistent within this field. Most studies evaluate ML/DL models against traditional ARIMA/ARIMA(X) techniques, with no consistent best performing model, with ARIMA(X) [17], LSTM [15], RFR each outperforming the other [24].

Exploring multiple techniques may therefore reveal better models for our data.

Although not aiming to produce the optimal forecasting model, our Reddit and Baseline models MAE from 1–14-day forecasts increased by 283% and 534% respectively, sharper error increases than related research. Kellner [15] found both ‘Twitter’ and baseline ARIMAX models MAE increased by 32%, and ‘Twitter’ and baseline LSTM models 39.25% and 6.16% respectively from 1-14-days. Kellner’s noticeably lower baseline LSTM MAE increase suggests alternative techniques would better capture temporal dependencies in our baseline model, impacting the improvement we report our Reddit features provide against baseline models.

## **5.5 Future Research**

Our discussion identified several avenues for future research, centred on improving quality, scope, and investigating alternative methodological approaches, defined below.

* **Quality:** Data collection via larger initial sets of Covid-specific keywords and improving class representation by extensive-keyword filtering may reveal new insights from data not collected by our dataset and improve classification performance by reducing class imbalance [20], potentially revealing Reddit as more predictive.
* **Scope:** Extending Reddits forecasting to other diseases and geographies would establish its utility within the broader disease surveillance field, assessing if Reddits forecasting utility is not just limited to diseases with sufficiently large public discussion or geographical representation, which this research cannot generalise.
* **Classification:** Exploring simpler ML techniques against complex models like RoBERTa would clarify the level of complexity our classification task requires, streamlining forecasting deployment, and improving classification performance and forecasting validity.
* **Regression:** Given our notable diminishing forecasting performance over increasing horizons, alternative models like LSTM may improve validity of our stated Reddit features utility for forecasting, whilst optimising performance would better contextualise Reddits utility against established resources.
* **Information-extraction:** Alternative methodologies may reveal alternative indicators of disease spread. Sentiment analysis could monitor adherence to public-health interventions for forecasting, and topic-modelling could reveal emerging topic/themes with predictive insight, unrelated to direct statements of infections [17]. This would validate Reddits’ utility in forecasting beyond our classification-based approach like established digital resources.

# **6 Conclusion**

Our research aimed to assess Reddits utility within disease forecasting, and the urgency of our investigation aligns with recent paywall restrictions on Twitter, leading to our exploration of Reddit as an alternative resource, potentially holding similar indicators to disease which have proved vital for disease surveillance systems [44].

## **6.1 Findings**

Our case study revealed Reddit as an informative resource for Covid-19 forecasting, providing direct and indirect indicators of disease prevalence and trajectory. We identified personal accounts of infection as consistently strong predictors for Covid-19 incidence, showing a strong correlation of 0.95. Our forecasting exercise then validated our Reddit features informativeness, improving baseline RFR MAE by 23.26% and 31.83% for 7 and 14-day forecasts respectively, producing more robust models with less predictive variation.

Contextualising within related literature revealed Reddits potential as an alternative resource, providing comparatively effective classification-based indicators for disease forecasting. However, at granular inspection, Reddit features provided little information to baseline models at a significant Covid-19 peak, questioning our features reliability as predictive indicators to disease.

**6.2 Methodological Insights**

A case study design was well suited for realising our aim and objectives, requiring an in-depth mixed-methods approach to exploration and adoption of proven methodologies to our unexplored resource.

BERT performed well for our classification task considering its complexity, and minimal text pre-processing requirements simplified this task. RFR likewise was easily utilised, and its transparency helped inform our overall assessment of our Reddit features predictive utility.

Our longitudinal forecasting exercise validated our features predictiveness beyond identifying relationships, allowing for model evaluation over various disease dynamics that a cross-sectional design would not capture.

## **6.3 Limitations**

Comparisons to research however revealed methodological limitations. Our reliance on single classification and regression models affects our findings validity. Exploring multiple techniques may result in improved model performance for both tasks, affecting the extent our Reddit features were found to be informative for forecasting Covid-19. Relatedly, the informativeness of suspected/ambiguous infection texts remains unclear. Its lower classification precision introduces noise, limiting its true predictive value, or confounding variables may hold predictive insights rather than these texts.

Additionally, our exploration of Reddit identified inherent challenges in its real-world application. Reddits lengthy texts demands considerable time in classification model training, limiting real-time insights and model deployment, whilst Reddits lack of geospatial data limits its forecasting application to specific geographies, a critical requirement for disease surveillance systems. Although our subreddit filtering approach can approximate geospatial information, these limitations starkly contrast the characteristics of established digital resources within this field.

## **6.4 Reflections**

Whilst our investigation finds Reddit to provide predictive indicators to Covid-19, revealing its potential within the wider disease forecasting field, its real-world utility faces distinct limitations and challenges.

However, continued restricted access to Twitter forces re-examination and exploration of novel alternative digital resources for disease surveillance like ours with Reddit, which can reveal comparable predictive insights.

Ultimately, our exploration of Reddit, and the continued exploration of new resources helps in the collective preparedness for future pandemics.

# **References**

[1] V. Lampos *et al.*, "Tracking COVID-19 using online search," *NPJ Digit Med,* vol. 4, no. 1, p. 17, Feb 8 2021, doi: 10.1038/s41746-021-00384-w.

[2] A. Gupta and R. Katarya, "Social media based surveillance systems for healthcare using machine learning: A systematic review," *J Biomed Inform,* vol. 108, p. 103500, Aug 2020, doi: 10.1016/j.jbi.2020.103500.

[3] R. Badker *et al.*, "Challenges in reported COVID-19 data: best practices and recommendations for future epidemics," *BMJ Glob Health,* vol. 6, no. 5, May 2021, doi: 10.1136/bmjgh-2021-005542.

[4] A. E. Aiello, A. Renson, and P. N. Zivich, "Social Media- and Internet-Based Disease Surveillance for Public Health," *Annu Rev Public Health,* vol. 41, pp. 101-118, Apr 2 2020, doi: 10.1146/annurev-publhealth-040119-094402.

[5] G. Chowell, L. Sattenspiel, S. Bansal, and C. Viboud, "Mathematical models to characterize early epidemic growth: A review," *Phys Life Rev,* vol. 18, pp. 66-97, Sep 2016, doi: 10.1016/j.plrev.2016.07.005.

[6] C. S. Lutz *et al.*, "Applying infectious disease forecasting to public health: a path forward using influenza forecasting examples," *BMC Public Health,* vol. 19, no. 1, p. 1659, Dec 10 2019, doi: 10.1186/s12889-019-7966-8.

[7] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature,* vol. 457, no. 7232, pp. 1012-4, Feb 19 2009, doi: 10.1038/nature07634.

[8] A. Culotta, "Towards detecting influenza epidemics by analyzing Twitter messages," in *1st Workshop on Social Media Analytics*, Washington, DC, USA, SOMA, Ed., 2010.

[9] S. V. Nuti *et al.*, "The use of google trends in health care research: a systematic review," *PLoS One,* vol. 9, no. 10, p. e109583, 2014, doi: 10.1371/journal.pone.0109583.

[10] S. Kandula, S. Pei, and J. Shaman, "Improved forecasts of influenza-associated hospitalization rates with Google Search Trends," *J R Soc Interface,* vol. 16, no. 155, p. 20190080, Jun 28 2019, doi: 10.1098/rsif.2019.0080.

[11] S. M. Ayyoubzadeh, S. M. Ayyoubzadeh, H. Zahedi, M. Ahmadi, and R. N. K. S, "Predicting COVID-19 Incidence Through Analysis of Google Trends Data in Iran: Data Mining and Deep Learning Pilot Study," *JMIR Public Health Surveill,* vol. 6, no. 2, p. e18828, Apr 14 2020, doi: 10.2196/18828.

[12] N. L. Bragazzi and N. Mahroum, "Google Trends Predicts Present and Future Plague Cases During the Plague Outbreak in Madagascar: Infodemiological Study," *JMIR Public Health Surveill,* vol. 5, no. 1, p. e13142, Mar 8 2019, doi: 10.2196/13142.

[13] A. Husnayain, A. Fuad, and L. Lazuardi, "Correlation between Google Trends on dengue fever and national surveillance report in Indonesia," *Glob Health Action,* vol. 12, no. 1, p. 1552652, 2019, doi: 10.1080/16549716.2018.1552652.

[14] C. A. Marques-Toledo *et al.*, "Dengue prediction by the web: Tweets are a useful tool for estimating and forecasting Dengue at country and city level," *PLoS Negl Trop Dis,* vol. 11, no. 7, p. e0005729, Jul 2017, doi: 10.1371/journal.pntd.0005729.

[15] D. Kellner, D. Lowin, and O. Hinz, "Improved healthcare disaster decision-making utilizing information extraction from complementary social media data during the COVID-19 pandemic," *Decision Support Systems,* 2023, doi: 10.1016/j.dss.2023.113983.

[16] L. Samaras, E. Garcia-Barriocanal, and M. A. Sicilia, "Comparing Social media and Google to detect and predict severe epidemics," *Sci Rep,* vol. 10, no. 1, p. 4747, Mar 16 2020, doi: 10.1038/s41598-020-61686-9.

[17] R. Lamsal, A. Harwood, and M. R. Read, "Twitter conversations predict the daily confirmed COVID-19 cases," *Appl Soft Comput,* vol. 129, p. 109603, Nov 2022, doi: 10.1016/j.asoc.2022.109603.

[18] F. Wang *et al.*, "Regional Level Influenza Study with Geo-Tagged Twitter Data," *J Med Syst,* vol. 40, no. 8, p. 189, Aug 2016, doi: 10.1007/s10916-016-0545-y.

[19] M. Salathe *et al.*, "Digital epidemiology," *PLoS Comput Biol,* vol. 8, no. 7, p. e1002616, 2012, doi: 10.1371/journal.pcbi.1002616.

[20] M. J. Paul, M. Dredze, and D. Broniatowski, "Twitter Improves Influenza Forecasting," *PLoS Currents,* 2014, doi: 10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117.

[21] M. Santillana, A. T. Nguyen, M. Dredze, M. J. Paul, E. O. Nsoesie, and J. S. Brownstein, "Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance," *PLoS Comput Biol,* vol. 11, no. 10, p. e1004513, Oct 2015, doi: 10.1371/journal.pcbi.1004513.

[22] M. Odlum and S. Yoon, "What can we learn about the Ebola outbreak from tweets?," *Am J Infect Control,* vol. 43, no. 6, pp. 563-71, Jun 2015, doi: 10.1016/j.ajic.2015.02.023.

[23] M. Abouzahra and J. Tan, "Twitter vs. Zika—The role of social media in epidemic outbreaks surveillance," *Health Policy and Technology,* vol. 10, no. 1, pp. 174-181, 2021, doi: 10.1016/j.hlpt.2020.10.014.

[24] C. Shen, A. Chen, C. Luo, J. Zhang, B. Feng, and W. Liao, "Using Reports of Symptoms and Diagnoses on Social Media to Predict COVID-19 Case Counts in Mainland China: Observational Infoveillance Study," *J Med Internet Res,* vol. 22, no. 5, p. e19421, May 28 2020, doi: 10.2196/19421.

[25] A. Sarker and Y. Ge, "Mining long-COVID symptoms from Reddit: characterizing post-COVID syndrome from patient reports," *JAMIA Open,* vol. 4, no. 3, p. ooab075, Jul 2021, doi: 10.1093/jamiaopen/ooab075.

[26] M. D. Choudhury and D. Sushovan, "Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity," in *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[27] G. Gkotsis *et al.*, "Characterisation of mental health conditions in social media using Informed Deep Learning," *Sci Rep,* vol. 7, p. 45141, Mar 22 2017, doi: 10.1038/srep45141.

[28] R. Keshavamurthy, S. Dixon, K. T. Pazdernik, and L. E. Charles, "Predicting infectious disease for biopreparedness and response: A systematic review of machine learning and deep learning approaches," *One Health,* vol. 15, p. 100439, Dec 2022, doi: 10.1016/j.onehlt.2022.100439.

[29] Y. Yang, S. F. Tsao, M. A. Basri, H. H. Chen, and Z. A. Butt, "Digital Disease Surveillance for Emerging Infectious Diseases: An Early Warning System Using the Internet and Social Media Data for COVID-19 Forecasting in Canada," *Stud Health Technol Inform,* vol. 302, pp. 861-865, May 18 2023, doi: 10.3233/SHTI230290.

[30] Y. Didi, A. Walha, M. Ben Halima, and A. Wali, "COVID-19 Outbreak Forecasting Based on Vaccine Rates and Tweets Classification," *Comput Intell Neurosci,* vol. 2022, p. 4535541, 2022, doi: 10.1155/2022/4535541.

[31] Y. Peng, C. Li, Y. Rong, C. P. Pang, X. Chen, and H. Chen, "Real-time Prediction of the Daily Incidence of COVID-19 in 215 Countries and Territories Using Machine Learning: Model Development and Validation," *J Med Internet Res,* vol. 23, no. 6, p. e24285, Jun 14 2021, doi: 10.2196/24285.

[32] D. M. Low, L. Rumker, T. Talkar, J. Torous, G. Cecchi, and S. S. Ghosh, "Natural Language Processing Reveals Vulnerable Mental Health Support Groups and Heightened Health Anxiety on Reddit During COVID-19: Observational Study," *J Med Internet Res,* vol. 22, no. 10, p. e22635, Oct 12 2020, doi: 10.2196/22635.

[33] C. A. Melton, O. A. Olusanya, N. Ammar, and A. Shaban-Nejad, "Public sentiment analysis and topic modeling regarding COVID-19 vaccines on the Reddit social media platform: A call to action for strengthening vaccine confidence," *J Infect Public Health,* vol. 14, no. 10, pp. 1505-1512, Oct 2021, doi: 10.1016/j.jiph.2021.08.010.

[34] P. Seroyizhko, Z. Zhexenova, M. Shafiq, F. Merizzi, A. Galassi, and F. Ruggerio, "A Sentiment and Emotion Annotated Dataset for Bitcoin Price Forecasting Based on Reddit Posts," in *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, Abu Dhabi, United Arab Emirates, 2022: Association for Computational Linguistics.

[35] M. Guo *et al.*, "Identifying COVID-19 cases and extracting patient reported symptoms from Reddit using natural language processing," *Sci Rep,* vol. 13, no. 1, p. 13721, Aug 22 2023, doi: 10.1038/s41598-023-39986-7.

[36] A. Signorini, A. M. Segre, and P. M. Polgreen, "The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic," *PLoS One,* vol. 6, no. 5, p. e19467, May 4 2011, doi: 10.1371/journal.pone.0019467.

[37] L. Chen, K. S. M. T. Hossain, P. Butler, N. Ramakrishnan, and B. A. Prakash, "Flu Gone Viral: Syndromic Surveillance of Flu on Twitter Using Temporal Topic Models," presented at the 2014 IEEE International Conference on Data Mining, 2014.

[38] J. D. Ford *et al.*, "Temporal trends in health worker social media communication during the COVID-19 pandemic," *Res Nurs Health,* vol. 45, no. 6, pp. 636-651, Dec 2022, doi: 10.1002/nur.22266.

[39] J. Santos and S. Matos, "Analysing Twitter and web queries for flu trend prediction," *Theoretical Biology and Medical Modelling,* vol. 11, no. 1, 2014.

[40] S. Das and A. K. Kolya, "Predicting the pandemic: sentiment evaluation and predictive analysis from large-scale tweets on Covid-19 by deep convolutional neural network," *Evol Intell,* vol. 15, no. 3, pp. 1913-1934, 2022, doi: 10.1007/s12065-021-00598-7.

[41] S. Mohan, A. K. Solanki, H. K. Taluja, Anuradha, and A. Singh, "Predicting the impact of the third wave of COVID-19 in India using hybrid statistical machine learning models: A time series forecasting and sentiment analysis approach," *Comput Biol Med,* vol. 144, p. 105354, May 2022, doi: 10.1016/j.compbiomed.2022.105354.

[42] T. Daghriri, M. Proctor, and S. Matthews, "Evolution of Select Epidemiological Modeling and the Rise of Population Sentiment Analysis: A Literature Review and COVID-19 Sentiment Illustration," *Int J Environ Res Public Health,* vol. 19, no. 6, Mar 9 2022, doi: 10.3390/ijerph19063230.

[43] S. M. E. Arimaki, M. Morita, R. Barzilay, "Twitter Catches The Flu: Detecting Influenza Epidemics using Twitter," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK., 2011: Association for Computational Linguistics.

[44] C. Garzo ́n-Alfonso and M. Rodr ́ıguez-Mart ́ınez, "Twitter Health Surveillance (THS) System," presented at the IEEE International Conference on Big Data (Big Data), 2018.

[45] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arxiv* vol. 1, 2019.

[46] S. Chadsuthi, C. Modchang, Y. Lenbury, S. Iamsirithaworn, and W. Triampo, "Modeling seasonal leptospirosis transmission and its association with rainfall and temperature in Thailand using time-series and ARIMAX analyses," *Asian Pac J Trop Med,* vol. 5, no. 7, pp. 539-46, Jul 2012, doi: 10.1016/S1995-7645(12)60095-9.

[47] A. I. Saba and A. H. Elsheikh, "Forecasting the prevalence of COVID-19 outbreak in Egypt using nonlinear autoregressive artificial neural networks," *Process Saf Environ Prot,* vol. 141, pp. 1-8, Sep 2020, doi: 10.1016/j.psep.2020.05.029.

[48] S. Singh, K. S. Parmar, S. J. S. Makkhan, J. Kaur, S. Peshoria, and J. Kumar, "Study of ARIMA and least square support vector machine (LS-SVM) models for the prediction of SARS-CoV-2 confirmed cases in the most affected countries," *Chaos Solitons Fractals,* vol. 139, p. 110086, Oct 2020, doi: 10.1016/j.chaos.2020.110086.

[49] S. Dixon, R. Keshavamurthy, D. H. Farber, A. Stevens, K. T. Pazdernik, and L. E. Charles, "A Comparison of Infectious Disease Forecasting Methods across Locations, Diseases, and Time," *Pathogens,* vol. 11, no. 2, Jan 29 2022, doi: 10.3390/pathogens11020185.

[50] M. J. Kane, N. Price, M. Scotch, and P. Rabinowitz, "Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks," *BMC Bioinformatics,* vol. 15, no. 276, 2014.

[51] F. Shahid, A. Zameer, and M. Muneeb, "Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM," *Chaos Solitons Fractals,* vol. 140, p. 110212, Nov 2020, doi: 10.1016/j.chaos.2020.110212.

[52] J. C. a. J. D. Creswell, *Research Design. Qualitative, Quantitative and Mixed Methods Approaches*, 5th ed. Los Angeles: SAGE Publications, , 2018.

[53] H. Tyralis and G. Papacharalampous, "Variable Selection in Time Series Forecasting Using Random Forests," *Algorithms,* vol. 10, no. 4, 2017, doi: 10.3390/a10040114.

[54] WHO. *COVID-19 case and death data*.

[55] SocialGrep. *The Reddit COVID Dataset*.

[56] A. Vasmani *et al.*, "Attention Is All You Need," in *31st Conference on Neural Information Processing Systems*, Long Beach, CA, USA., 2017: NIPS 2017.

[57] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to Fine-Tune BERT for Text Classification?," in *China National Conference on Chinese Computational Linguistics*, Kunming, China, t. C. N. Conference, Ed., 2019: Chinese Computational Linguistics, pp. 194-206.

[58] Y. Liu *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692,* 2019.

[59] L. Breiman, "Random Forests," *Machine Learning,* vol. 45, pp. 5–32, 2001.

[60] D. R. Karp *et al.*, "Ethical and practical issues associated with aggregating databases," *PLoS Med,* vol. 5, no. 9, p. e190, Sep 23 2008, doi: 10.1371/journal.pmed.0050190.

[61] E. Aramaki, S. Maskawa, and M. Morita, "Twitter Catches The Flu: Detecting Influenza Epidemics using Twitter," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK, 2011: Association for Computational Linguistics, pp. 1568–1576.

[62] V. Lampos, T. De Bie, and N. Cristianini, "Flu Detector - Tracking Epidemics on Twitter," in *2nd International Workshop on Cognitive Information Processing, 2010*, 2010, pp. 411–416.

[63] S. Chatterjee, K. Ghosh, A. Banerjee, and S. Banerjee, "Forecasting COVID-19 Outbreak Through Fusion of Internet Search, Social Media, and Air Quality Data: A Retrospective Study in Indian Context," *IEEE Transactions on Computational Social Systems,* vol. 10, no. 3, pp. 1017-1028, 2023, doi: 10.1109/tcss.2022.3140320.

[64] S. C. Hogue *et al.*, "Pharmacists' perceptions of a machine learning model for the identification of atypical medication orders," *J Am Med Inform Assoc,* vol. 28, no. 8, pp. 1712-1718, Jul 30 2021, doi: 10.1093/jamia/ocab071.

# **Appendices**

## **Appendix A**

**Time Series Features and Descriptions**

|  |  |
| --- | --- |
| **Dataset Feature** | **Description** |
| Label: ('*YES'/'MAYBE'/'CONTACT'*) | No. Reddit texts classified as labels;  Author: 'Yes' / 'Maybe', or Contact: 'Yes', grouped by day |
| {*Label*}\_Rolling\_Mean | Rolling average (n=7) of respective label |
| [64]\_std | Rolling standard deviation (n=7) of respective label |
| {*Label}\_*diff | Rate of change between current and previous label value |
| {*Label*}\_Rolling\_Mean\_dfff | Rate of change between current and previous label (mean) value |
| {*Label*}\_std\_diff | Rate of change between current and previous label std |
| New\_cases | WHO reported Covid-19 incidence (Target Variable) |
| Previous\_New\_cases | Covid-19 incidence shifted by 1-day |
| 7\_day\_New\_cases | Covid-19 incidence lagged by 7 days (Target Variable) |
| 14\_day\_New\_cases | Covid-19 incidence lagged by 14 days (Target Variable) |

**Figure 1:** Summarising Our generated time-series features and engineered features, including our lagged features. {Label} represents and of the label ‘YES’, ‘MAYBE’, ‘CONTACT’, stating that for each label we produced rolling means, standard deviations, differencing etc.

Features ‘New\_cases’ ‘7\_day\_New\_cases’, ‘14\_day\_New\_cases, represent our targert features dependent on the forecasting horizon investigated.

## **Appendix B**

**BERT Training, Optimisation, Validation**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | **Multi- classification** | | | **Binary-classification** | |
|  |  | Mean | std | Mean | | std |
| **YES** | Precision | 0.76 | 0.03 | 0.74 | | 0.06 |
|  | Recall | 0.74 | 0.05 | 0.76 | | 0.04 |
|  | F1 | 0.75 | 0.02 | 0.75 | | 0.02 |
| **MAYBE** | Precision | 0.62 | 0.02 | 0.6 | | 0.11 |
|  | Recall | 0.64 | 0.05 | 0.63 | | 0.12 |
|  | F1 | 0.63 | 0.02 | 0.59 | | 0.04 |

Table 1: ‘Author’ dimension labels classification performance when classified as a multi-class classification task or individual binary classification tasks.

***A graph of different colored bars

Description automatically generated with medium confidence***

Figure 1: Mean F1-scores classification performance for each label when returning model based on different early stopping parameters; validation loss, and validation accuracy.

A graph of different levels of scores

Description automatically generated with medium confidence

Figure 2: Mean F1-score plots for each classification label, investigating three different learning rates for each classification model.

A graph of different scores

Description automatically generated with low confidence

Figure 3:Mean F1-score plots for each classification label, investigating different dropout rates for each classification model.

A graph of different types of lines

Description automatically generated with medium confidence

Figure 4:Mean F1-score plots for each classification label, investigating different BERT classification head architectures for each classification model.

**A graph of different colored bars

Description automatically generated**

Figure 5: Mean F1-score classification performance for Text with stemming and lemmetisation applied (NLP processed) vs our raw non-NLP text.

## **Appendix C**

**Correlational and Predictive visualisation**

A graph of a number of days lag

Description automatically generated

Figure 1:Spearman Lag correlations over a 28-day period of our non-smoothed and smoothed (rolling 7-day mean) classification label features against the Covid-19 incidence.

A graph showing the number of p values

Description automatically generated

Figure 2: Granger Causality test for our smoothed classification label features to the Covid-19 incidence, over a 28-day lag period, with the significance value p=0.05 plotted to aid visualisation.

## **Appendix D**

**Feature Importance Scores**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Forecasting Horizon** | | |
|  | **1 day** | **7 day** | **14 day** |
|  |  |  |  |
| **Previous New cases** | 0.52454 | 0.43789 | 0.22199 |
| **Yes rolling mean** | 0.33367 | 0.32013 | 0.40536 |
| **Contact rolling mean** | 0.03244 | 0.0637 | 0.08157 |
| **Maybe rolling mean** | 0.04401 | 0.09054 | 0.18825 |
| **Yes STD** | 0.01381 | 0.0264 | 0.03121 |
| **Maybe STD** | 0.01757 | 0.02377 | 0.02133 |
| **Contact STD** | 0.00932 | 0.01544 | 0.02611 |
| **Yes STD Diff** | 0.00274 | 0.00276 | 0.00411 |
| **Contact STD Diff** | 0.00726 | 0.00446 | 0.00294 |
| **Maybe STD Diff** | 0.00776 | 0.00446 | 0.00451 |
| **Yes rolling mean Diff** | 0.00176 | 0.00297 | 0.003 |
| **Maybe rolling mean Diff** | 0.00227 | 0.00339 | 0.0048 |
| **Contact rolling mean Diff** | 0.00287 | 0.00409 | 0.00482 |

Table 1: Average feature importance scores for each features within our time-series dataset, averaged over all testing periods relative to that forecasting horizon.

**Appendix E**

**Forecasting Performance at individual testing period**

A screenshot of a report

Description automatically generated

Table 1: Individual RFR performance for both baseline and Reddit models at individual testing periods for each forecasting horizon.

## **Appendix F**

**Artifacts directory and descriptions**

The following artifacts are provided in the complementary zip file, with descriptions and file names below:

Original Research Proposal submitted:

* ‘ResearchProposalEdited.pdf’

Ethics fast-track application:

* ‘CS Fast-Track ethical Approval Form\_Nathanael\_WASigned.pdf’

Ethics fast-track application approval screenshot:

* ‘EthicsApproval.docx’

Full list of USA and Covid-19 keywords used for subreddit filtering:

* ‘USAplusCOVIDkeywordsForSubredditFiltering .docx’

The pre-processed and filtered Reddit dataset:

* ‘USA\_and\_Covid\_specific\_Dataset.csv’

Our Multi-classification BERT model code in Google Colab:

* ‘MultiClassBERTcode.docx’

Our Binary-classification BERT model code in Google Colab:

* ‘BinaryClassBERTcode.docx’

Our Multi-classification BERT model:

* ‘MODEL\_User\_final\_2e\_d02\_ep5’ folder

Our Binary-classification BERT model:

* ‘MODEL\_Contact\_final\_2e\_d02\_ep4’ folder

Links to BERT models if issues with provided BERT models:

* ‘BERTmodelsLink.docx’

The generated time series dataset after BERT classification of above dataset, all features except target features are normalised:

* ‘Time\_Series\_Reddit\_classified\_Dataset.csv’

Brief information regarding our time-series datasets, what each feature represents:

* ‘DatasetInformation.docx’

Code snippets showing our training and test set generation, feature validation procedure, RFR hyperparameter tuning, Walk-Forward testing, Average metric generation:

* ‘ForecastingExercise.docx’

Visualisation/guide to help explain our Methodology:

* Methodology Overview.docx