Research Proposal

Department of Computer Science
COM00150M

Title:

Investigating the Covid-19 Forecasting Capability of Covid-related Reddit posts using NLP and BERT.

Table of Contents

1 Focus of Research

1.1 Field of study

Digital Epidemiology [1].

1.2 Topic/Research object

My research topic is Digital Disease forecasting, focusing on investigating Reddit's capability for Covid-19 forecasting.

Research Object(s):

- Reddit Discussions relating to Covid
- Historical Covid-19 data.
- LSTM model accuracy.

1.3 Unit of analysis

- Reddit – Reddit textual data (Posts/Comments) relating to Covid in the USA.
- Covid-19 - daily Covid-19 incidence (Cases) in USA.
- LSTM– Covid incidence forecasting accuracy.

1.4 Justification of the project – why is it important?

This project is suitable as an IRP at Masters level, demonstrating knowledge of advanced Deep-learning techniques, integration/pre-processing of various data-sources underscoring problem-solving skills, and the application of known methods to new problems. It therefore meets the complexity required for an IRP and has a required focus on Data Analytics.

This project additionally is relevant due to the Covid-19 Pandemic world-wide impact and recentness, making this topic well-known, and it is interesting and unique as it assesses Reddit's utility in Disease forecasting which few studies have explored.

1.5 Project Aim

The central aim is to determine if Reddit textual data can be an effective source for Covid-19 forecasting, with a comparison against traditional data sources alone.

The Questions and objectives below allow this to be determinable.

1.6 Main research question and supportive research questions and objectives

Research Question:

Can classification of Covid-19 related discussions on Reddit using NLP techniques and BERT be used to improve LSTM Covid-19 forecasting model accuracy compared to traditional data alone?

Supporting Research Questions:

1. Can NLP techniques and BERT accurately classify Covid related Reddit Posts/discussions into a 'User Dimension' and a 'Contact dimension'

1. Do the NLP derived classifications of Covid related Reddit Discussions correlate with Covid-19 incidence in the USA from March – December 2020?

2. Can LSTM models developed from Reddit data, and traditional metrics accurately predict the incidence of Covid-19 in the USA from March – December 2020?

3. How does integration of Reddit symptom-related discussions improve or alter the forecasting accuracy of LSTM models when compared to models based solely on traditional metrics and Google data?

Supporting Objectives:

1. Review literature on digital disease forecasting with a focus on LSTM.
2. Acquire, pre-process and classify covid-related Reddit data using BERT.
3. Acquire, pre-process historical Covid-incidence (traditional metrics).
4. Construct and fine-tune LSTM models, ensuring their validity.
5. Quantitatively compare predictive accuracies of the LSTM models.
6. Dissect findings, identify strengths, limitations, and future prospects.

1.7 Research Question Justification

The main research question specifies this study's goal, while supportive questions provide:

1. Preliminary evidence for relationships between the BERT derived classes and Covid-19 cases.
2. Address forecasting utility of Reddit data
3. Assess added value of Reddit in forecasting models,

Altogether these effectively answer the main research question quantitatively. Whilst additional objectives help deconstruct the problem in a structured manner.

The questions limit geographic and temporal scope to improve the research's quality and mitigate unnecessary complexity.

The research questions focus on measuring/comparing forecasting model accuracy is in-line with related research [1], and therefore suitable to realise this projects aim.

2 Literature Review

2.1 Introduction

Disease forecasting plays a crucial role in preparedness and countermeasures against infectious diseases [2] with its importance highlighted by the Covid-19 pandemic. Though traditionally utilising structured clinical data, modern Digital epidemiology has redefined disease forecasting, leveraging large-scale real-time internet-based data and advanced forecasting methods to improve disease forecasting accuracy [3]. This ultimately helped inform government action to Covid-19 [4] and emphasises internet-based data's and modern forecasting techniques importance within epidemiology.

This literature review focuses on internet-based data within disease surveillance, the advancement of forecasting methods and justification of Reddit being a potential un-investigated source for forecasting and the mechanism for its assessment

2.2 Traditional forecasting

Disease forecasting, subfield of epidemiology, traditionally employed structured surveillance/clinical data, and statistical methods like time-series forecasting and compartmental approaches like SIR models [2]. This traditional methodology helped in response to outbreaks like Spanish Flu and AIDS [2]. However, a limitation to this approach is high latency from data collection to actions, causing response lags up to two weeks [5]. Badker [6] concluded such delays affected timely countermeasures to Covid-19, stressing the importance of timely insights.
Modern pandemics like Covid-19 have re-highlighted the urgency for real-time, insightful data, and Digital epidemiology has explored the integration of internet-driven data sources for more timely and accurate forecasting [3].

2.3 Internet-based data in forecasting

At the forefront of Digital epidemiology was Google data, and Ginsberg's study [7] was a major milestone, accurately predicting Influenza incidence in real-time using Google search data, doing so two weeks ahead of CDC reports. This lead to development of 'Google Flu Trends' [8], and highlighted the improved timeliness and accuracy internet-based data provides to forecasting.
By 2014, Nuti's review of disease surveillance identified 70 studies incorporating Google Trends [9], highlighted Google's impact on the field, and has additionally enhanced forecasts for diseases like Dengue and Ebola [3].

Subsequently, User-Generated Content opened new opportunities. Social media data, notably Twitter, allowed researchers to capture public sentiments [10]. Twitters promise in disease forecasting however became apparent following appearing as a source in most influenza forecasting models submitted in 'FluSight' competitions [5]. Gupta's 2020 review [11] highlights Twitters dominance, revealing 64% of digital disease surveillance studies relied on Twitter. Twitters short texts and metadata alongside sentiment analysis and Natural Language Processing, provided new opportunities in disease surveillance [12, 13]. Highlighted by Wang [14] who improved regional influenza forecasting using Twitter geolocation metadata.

Despite extensive research inherent biases and limitations like data non-independence and representativeness [3], affect research validity.
Lampos's [4] study investigating Google predictive ability within Covid forecasting demonstrated confounding effect as media attention artificially surges Google search-terms, unrelated to disease prevalence. Additionally, Szmuda [15] identified that Google search trends for Covid-19 correlated more with WHO announcements than real-time Covid-19 spread, questioning its utility altogether. These limitations affect real-time applications, highlighted in 2014 when 'Google FluTrends' was discontinued due to significantly overestimating Influenza incidence during 2012-13 winter [16].

UGC suffers similar challenges of self-perpetuating trends [17], whilst limited user demographics to each platform may underrepresent vulnerable populations and subsequent developed models[3]. Whilst Twitters API, offering only a subset of tweets, potentially introduces sampling bias [3].
Awareness of these limitations is crucial to determine validity of research outcomes.

2.4 New Forecasting methods

Historically, disease forecasting relied on statistical models like ARIMA [2]. However, their linear assumptions struggle with non-linear dynamics of diseases and high-dimensionality of modern internet-based datasets [18].

Digital epidemiology therefore transitioned to Machine Learning (ML) and Deep Learning (DL) [11]. These techniques manage vast datasets and emphasise pattern recognition over statistical approaches, and as highlighted by Dixon [18] often outperform traditional methods in disease forecasting accuracy.

DL techniques like Long Short-Term Memory (LSTM) excel at sequence prediction and were heavily utilised for Covid-19 pandemic forecasting. Various studies [19-21] developed LSTM models from integrated datasets of traditional and internet-based data, to improve forecasting accuracy, highlighting LSTM's versatility in pandemic response.

Equally, Random-Forest-Regression (RFR) demonstrated utility during Covid-19. Peng et al. [22] combined Google with traditional data to accurately forecast 215 countries Covid-incidence 14 days in advance. Although RFR outperformed LSTM models in this study, inadequate training period was provided for LSTM. Kellner [13] however found LSTM models outperformed RFR in forecasting, further highlighted the hidden insights LSTM offers over ML models.

2.5 Investigating Reddit

Digital epidemiology now faces challenges. Twitter’s new API paywall restrictions presents a major setback to this field, and a re-examination of alternative data within epidemiology is needed in preparation for future pandemics.

Reddit emerges as an interesting alternative. Reddit’s in-depth discussions provides opportunities for epidemiological exploration through sentiment and textual analysis, generating insights into public behaviours [23]. While Reddit has been utilised for sentiment analysis, symptom and topic modelling, [24] its potential in disease forecasting and projection is limited.

Of these few such studies, Kellner et al. [13] demonstrated that the addition of features derived from Covid symptom discussion on Reddit can improve the accuracy of covid case and hospitalisation forecasts compared to text-classification of Covid-related tweets alone.
Kellner generated Covid-related symptom frequency indices from Reddit discussions using a lexicon-based approach, based on previous work by Sarker et al. [25]. This approach is consistent with other digital epidemiological research utilising Twitter and Google data via a function of topic frequency and search-term frequency respectively [4, 5, 7, 12, 13], and highlights Reddit’s potential in pandemic forecasting, however it stops short of evaluating Reddit's standalone efficacy in forecasting, which warrants further investigation.

A critical consideration of text analysis methods and approaches is needed before investigating Reddit's utility in COVID-19 forecasting. Kellner's lexicon-based symptom frequency method, although chosen due to being complimentary to the Twitter data, may not optimally exploit Reddit's data for predictive features, and methods like classification, sentiment analysis and topic modelling may be more informative. Additionally, this lexicon-based approach, while expanded and refined from Sarker et al. [25], could fall short in grasping linguistic subtleties and ensuring the relevance of symptom discussions to COVID-19, thus questioning the reliability of the derived features.

Building on the work by Sarker [25] , Guo et al. [26] employed a BERT-based model to first isolate COVID-specific Reddit discussions before extracting symptoms using an additional BioBERT model, ensuring a higher accuracy of COVID-related symptom identification. This method, validated through cross-verification and showing significant differences in Covid symptom frequency when compared to Sarker’s lexicon-based approach, suggests that advanced NLP techniques like BERT offer a more nuanced and reliable means of feature extraction from Reddit for use in forecasting models.

Based on Kellner's findings with Twitter, applying a similar classification framework to Reddit discussions could yield robust predictive features for Covid forecasting. Categorising Reddit texts into ‘User’ ('Has Covid' or 'May have Covid') and ‘Contact’ ('Family or colleague has Covid') dimensions, like Kellner et al [13], aligns with proven digital epidemiological research
Additionally, BERT's advanced NLP capabilities for classification tasks provides a compelling approach. This strategy, informed by Guo's [26] use of BERT for precise text classification and Kellner's effective use of user and contact categories, offers a

promising avenue for extracting meaningful indices from Reddit for disease forecasting models.

2.6 Assessing Utility in Forecasting

Correlational studies are frequent in disease forecasting, highlighting relationships between internet-data and disease metrics. For instance, Lampos et al. [4] used this approach to gauge the predictive potential of Google-search data for Covid metrics. Also frequent are comparison studies, evaluating the performance of different forecasting models to determine model or data-source utility. Kellner [13] used a comparative approach to demonstrate the enhanced forecasting ability when integrating Reddit with Twitter data. Other works, like [18] again compare accuracy between forecasting models across various diseases to determine ML and DL superiority to traditional methods.

In evaluating Reddit's utility for Covid-19 forecasting, correlation analysis should be performed between the derived Reddit text classifications and Covid metrics, providing preliminary predicative evidence, demonstrating relationships. Then, comparisons between forecasting models accuracy demonstrate the predictive potential in forecasting. This comparative approach assesses Reddit's viability over traditional data.

3 Research Methodology

3.1 Research Design and Methodology

From the literature review, disease forecasting using internet-data is commonly a function of topic or term frequency when using Twitter and Google data [4, 5, 7, 12, 13], but more advanced Natural Language Processing techniques are available with textual data. Based on Kellner's [13] and Guo's [26] methodology, this project proposes a Transformer based approach using BERT to classify Covid-related texts from Reddit discussions into a User or Contact dimension (see Section 3.2), generating daily indices as features for forecasting, in line with relevant literature. These classification indices features will be combined with daily Covid-case numbers in LSTM models. This forecasting accuracy will be compared to LSTM models derived using only traditional data (Covid-cases) as features.

The project is a longitudinal study, employing a correlational and comparative research design, allowing for comprehensive analysis of relationships between Reddit text classification indices and Covid-19 incidence, and assessing Reddit's forecasting utility via LSTM model comparison.

The correlational aspect is typical in Digital forecasting research [4] assessing strengths and direction of relationships between the BERT derived classes from Reddit discussions and daily Covid-case numbers provides preliminary predictive value evidence.
The comparative aspect compares the forecasting accuracy of different LSTM models. This comparison evaluates Reddit's utility as a forecasting source in relation to Traditional data alone, and contextualises the models results predictions.

A longitudinal design, as opposed to cross-sectional, keeps in-line with related literature [13, 22], enabling examination over extended periods, increasing LSTM model reliability, and evaluating forecasting performance across different phases of the pandemic.

3.2 Sampling

Reddit Data – Secondary data collected from 'SocialGrep' will be used as a convenience sample for this research. This dataset consists of 17 million Reddit posts/discussions that mention the word 'Covid' before 25.10.2021, and includes features that can isolate specific subreddits and group posts by time.

Historical Covid-19 Data – Convenience sampling of historical Covid-19 data previously collected/aggregated into datasets, available for researchers.

3.3 Data Collection and Pre-processing

Reddit Data:

Collection:
The Reddit 'Covid' dataset collected by 'SocialGrep' can be retrieved from the website: https://socialgrep.com/datasets/the-reddit-covid-dataset, in the form of two CSV files.

Pre-processing:
Pre-processing steps are focused on retrieving only relevant subreddits, cleaning and preparing textual data for BERT models to classify.

- Select subreddits indicative to the USA.
- Clean text from noise, URLs, and normalise.
- Tokenise text and subword tokenisation.
- Use padding and truncation so text is suitable length for BERT.
- Fine-tune BERT model with labelled sample.
- Classify Reddit texts into either a 'User' dimension:
  1. User Confirmed Covid Status
  2. User Suspected Covid Status
  3. No Covid Status

  Or 'Contact' dimension;
  1. Work Colleague confirmed status
  2. Family Member confirmed status
  3. 'Other' contact confirmed status

Historical Covid-19 Data:

Collection: Collected via WHO website: https://covid19.who.int/data
Data filtered by time and country.

Pre-processing:
Exploration and correction of missing values and anomalies.

3.4 Model Development and Evaluation

The study will develop LSTM models using different combinations of data: Covid-19 incidence alone as a baseline and Covid-19 incidence integrated with Reddit data classified using BERT. LSTM models are chosen due to their effectiveness in disease forecasting and ability to capture complex patterns [18].

For training and validation, a rolling-window approach will be utilized, forecasting Covid-19 incidence 14 days in advance. This approach enables hyperparameter tuning. The choice of 14-day forecasting period is supported by previous research demonstrating accurate forecasts within this timeframe [13, 22].
The rolling window approach will be applied to the test sets, allowing generation of Covid-19 incidence forecasts on unseen data.

Model accuracy will be evaluated on performance metrics, including Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), aligning with previous research [18].

Summary statistics, tables and graphs will aid the evaluation. Insights from this phase will directly address the study's primary research questions.

3.5 Critique and Limitations of the Methodology

This methodology addresses the research questions, utilising a typical approach in Digital epidemiological to assess a data source and/or forecasting techniques utility [13, 18]. However, limitations exist.

Geographic and temporal constraints limit scope to focus on USA between March–December 2020, limiting generalisability to other countries and timeframes [28]. This limited scope however ensures research quality, concentrating analysis and limiting complexity. An expanded geographic and temporal scope would require significant time and skill, beyond the requirement of this Independent-Research-Project.

Reliability issues may arise from failing/inaccurately detecting Covid-related symptoms from Reddit data, whilst the legitimacy of users' symptoms is uncertain [11].
Reliability issues may arise from failing/inaccurately classifying Reddit text using BERT models, whilst the legitimacy of users' statement cannot be verified [11]. Fine-tuning pre-trained BERT models improves textual classification accuracy, but relies in a labelled samples, of which some classes may be under-represented for in the data [27].

Reliability issues also extend to LSTM models. Inadequate care or experience to prevent overfitting or suboptimal-tuning may affect outcomes that do not reliably reflect the utility of Reddit in Covid-19 forecasting [19].

Additionally, this methodology assesses Reddit's utility for forecasting based on the classification of text into a User and Contact dimension. If results conclude Reddit derived models do not improve forecasting accuracy, it does not dismiss Reddit's utility if other methodologies were explored. Therefore, this approach cannot conclude Reddit's lack of utility outside this classification approach.

Limitations will be thoroughly discussed therefore contextualising the interpretation of findings, and improving validity by highlighting constraints [28]. And as is expected in Digital epidemiological research, limitations, biases and ethical implications inherent to internet-based data will be discussed [3].

4 Project Management

This section presents project management techniques, Work-Breakdown-Structure (WBS), Gantt Chart, Project Schedule (Timeline), Activity Network, to aid project organisation and realistic within 16-week period, whilst addressing limitations and resources.

4.1 Work Breakdown Structure (WBS)

WBS organises my project into hierarchical tasks, ensuring the entire scope is visualized and nothing missed, as seen in Image 1.

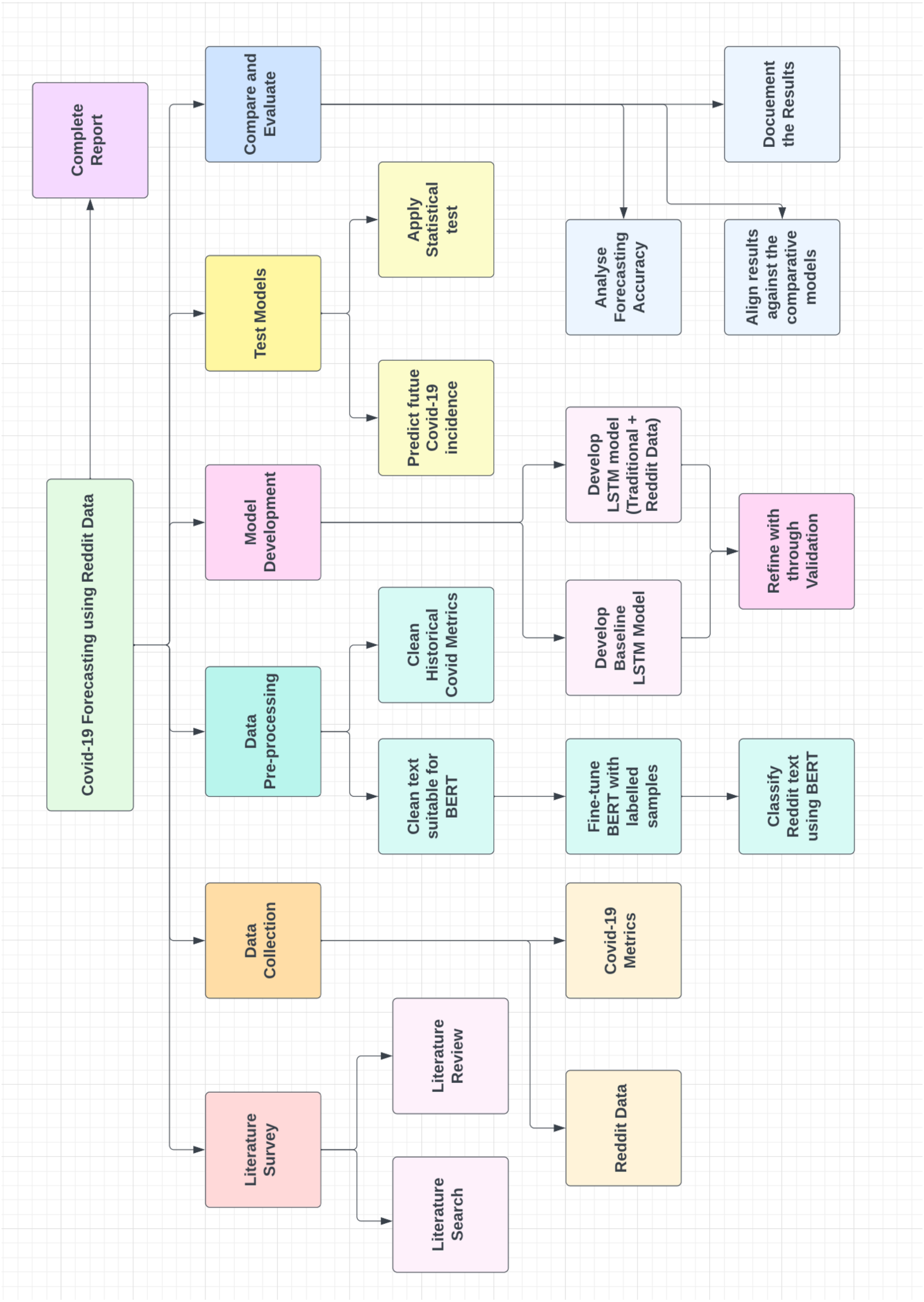
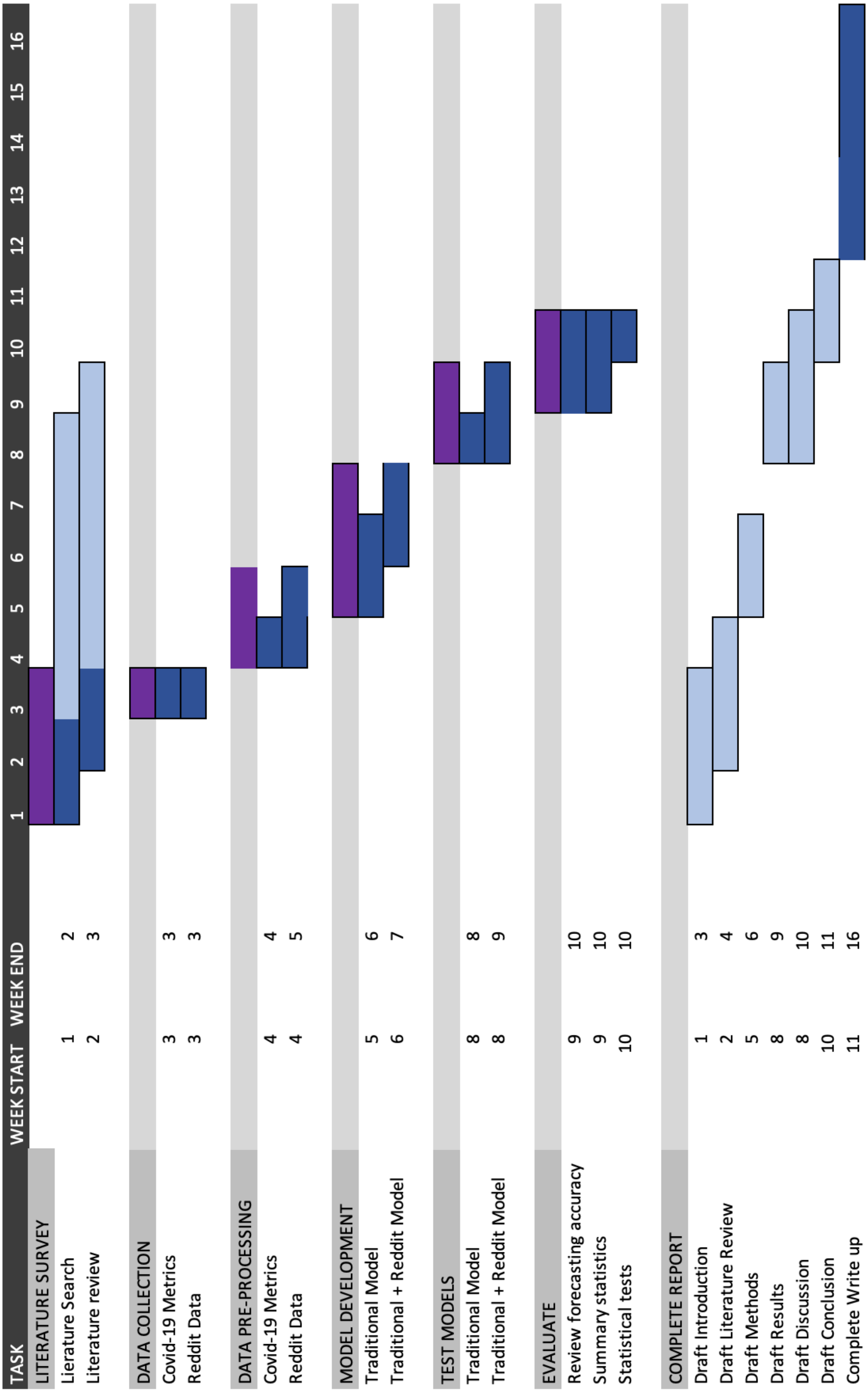

Covid-19 Forecasting using Reddit Data
Complete Report
Literature Survey
Literature Search
Literature Review
Data Collection
Reddit Data
Covid-19 Metrics
Data Pre-processing
Clean text suitable for BERT
Fine-tune BERT with labelled samples
Classify Reddit text using BERT
Clean Historical Covid Metrics
Model Development
Develop Baseline LSTM Model
Develop LSTM model (Traditional + Reddit Data)
Refine with through Validation
Test Models
Predict futue Covid-19 incidence
Apply Statistical test
Compare and Evaluate
Analyse Forecasting Accuracy
Align results against the comparative models
Docuement the Results

Image 1. WBS showing hierarchy of tasks.

4.2 Gantt Chart

The below Gantt chart derived from the WBS, shows allocated times for each task and provides visualisation of the project timeline, discussed following this. (Light blue indicates some time dedicated during this period)



TASK	WEEK START	WEEK END
LITERATURE SURVEY		
Lierature Search	1	2
Literature review	2	3
DATA COLLECTION		
Covid-19 Metrics	3	3
Reddit Data	3	3
DATA PRE-PROCESSING		
Covid-19 Metrics	4	4
Reddit Data	4	5
MODEL DEVELOPMENT		
Traditional Model	5	6
Traditional + Reddit Model	6	7
TEST MODELS		
Traditional Model	8	8
Traditional + Reddit Model	8	9
EVALUATE		
Review forecasting accuracy	9	10
Summary statistics	9	10
Statistical tests	10	10
COMPLETE REPORT		
Draft Introduction	1	3
Draft Literature Review	2	4
Draft Methods	5	6
Draft Results	8	9
Draft Discussion	8	10
Draft Conclusion	10	11
Complete Write up	11	16

Image 2. Gantt chart of Project plan and individual tasks.

4.3 Project Plan (Timeline)

Individual tasks are grouped hierarchically, identified by WBS, into ‘Phases. Each ‘Phase’ includes encompassing tasks, timing, dependencies, and risks. Specific time allocated to tasks within Phases shown in Gantt chart.

Phase 1: Literature Survey (Weeks 1-3)
Tasks:

- Initial collection of research relevant to my project.
- Begin drafting Introduction and Literature Review.

Duration: 3 weeks

Dependencies: None.

Potential Risks: Inclusion of material outside projects scope.

Phase 2: Data Collection (Week 3)
Tasks:

- Download Covid-19 metrics.
- Download Reddit dataset from ‘SocialGrep’ website.

Duration: 1 week

Dependencies: Ethical approval for collecting Reddit data.

Potential Risks:

- Delays in ethical approval
- Ethical considerations and restrictions accessing user-generated content on Reddit.

Phase 3: Data Pre-Processing (Weeks 4-5)
Tasks:

- Clean Covid-19 metric data
- Clean and pre-process Reddit data suitable for BERT
- Fine-tune BERT
- Classify Reddit text using BERT
- Correlation and Regression

Duration: 2 weeks

Dependencies: Successful data collection.

Potential Risks:

- Issues with data quality leading to extended cleaning and processing periods.

- Inadequate experience using NLP techniques and BERT.

Phase 4: Model Development (Weeks 5-7)
Tasks:

- Develop LSTM models on respective data.
- Training and Validation using a rolling-window approach.
- (Refinement and troubleshooting)

Duration: 3 weeks

Dependencies: Completed pre-processed datasets.

Potential Risks:

- Challenges with model complexity or performance.
- Need for extended tuning or knowledge beyond initial estimates.
- Inadequate experience using LSTM.

Phase 5: Test Models (Weeks 8-9)
Tasks:

- Test models on unseen test data.
- (Refinement and troubleshooting.)

Duration: 2 weeks

Dependencies: Trained and Validated LSTM models from Phase 4.

Potential Risks:

- Model underperformance on test data leading to more extended refinement periods.
- Inadequate experience using LSTM.

Phase 6: Evaluation (Weeks 9-10)
Tasks:

- Evaluate forecasting predictions.
- Generate descriptive statistics.
- Apply statistical tests.

Duration: 2 weeks

Dependencies: Tested models from Phase 5.

Potential Risks:

- Unforeseen discrepancies in model performance.

- Additional time requirements for thorough statistical analyses.

Phase 7: Complete Write-up (Weeks 11-16)
Tasks:

- Draft write-up ongoing before this point.
- Dedicated time for write-up and finalization.

Duration: 6 weeks

Dependencies: Completion of all prior phases.

Potential Risks:

- Delays in previous phases leading to reduced writing periods.
- Extended time required in previous phases.

4.4 Activity Network

Activity Networks highlights the sequences and dependencies between tasks, and identify Milestones in the project.

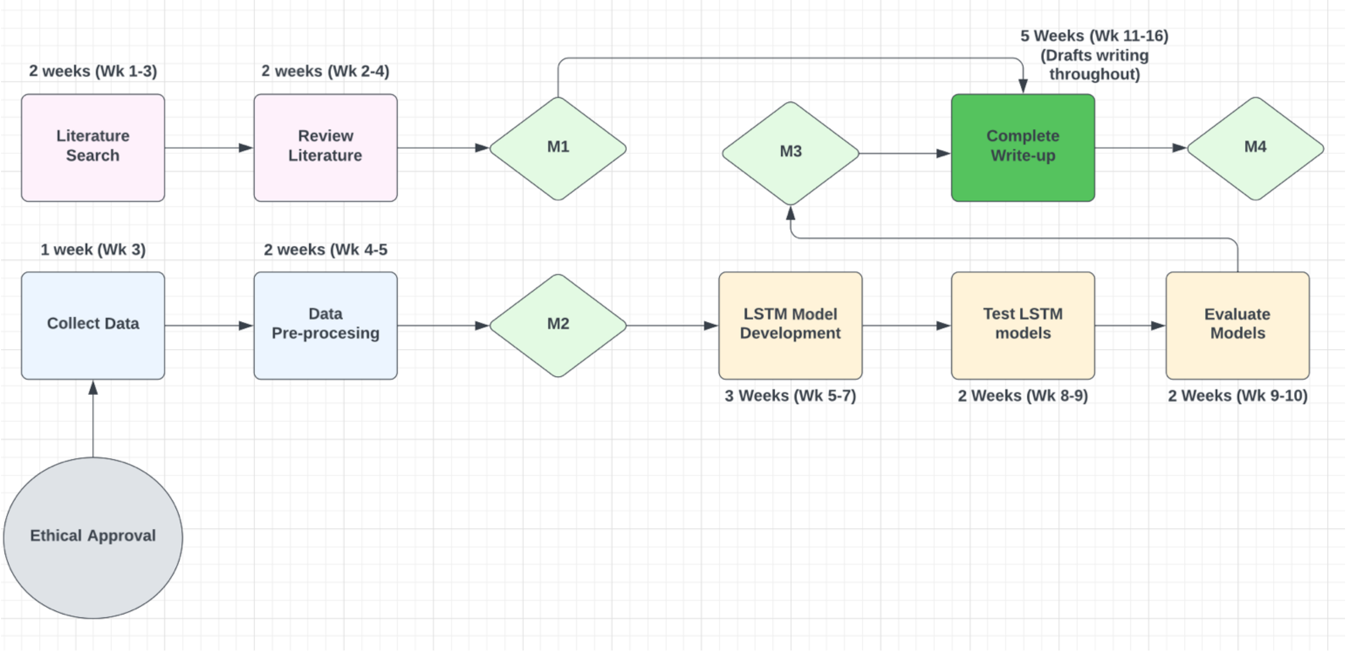


Image 3. Activity Network of project.

4.5 Milestones

Milestone 1 – Development of a Theoretical Foundation (Week 3):

- Corresponds to the completion of the Literature Survey.

Milestone 2 - Transition from Information Gathering to Model Generation (Week 5):

- This is after the Data Pre-Processing phase.

Milestone 3 - Data to Answer Research Questions (Week 9):

- Post the Model Testing phase.

Milestone 4 - Project Conclusion (Week 16):

- Following the Complete Write-up phase.

4.6 Resources and limitations.

Hardware:

Potential issues due to computational demand of BERT and LSTM models. Utilising Google Colab minimises this issue, providing increased computing power and GPU access.

Software:
Python, with libraries like Keras and TensorFlow, will be essential. Unexpected issues accessing libraries may arise on my personal computer.
These resources however are available on the Google Colab and mitigate this.

3. Skill set:
Acquiring skills in Natural Language Processing techniques, fine-tuning BERT models, and LSTM model development is crucial. Potential delays if these skills are not learnt beforehand. Additional time is allocated in respective phase timeline for unexpected issues.

4. Time Management:
Adhering to the time allocation for each task is essential, especially considering built-in buffer periods.

Risks are additionally highlighted within each ‘Phase’ of the timeline. Additionally, I've allocated more time to tasks I anticipate to be challenging, as mentioned above.

5 Ethical implication of this study

Developing disease forecasting models using user-generated content, introduces ethical implications.

Consent
Though social-media data is public, implied consent from users posting content publicly may not extend to academic purposes, raising the issue between implied vs explicit consent [28]. Though a well referenced issue, my aim is to mitigate this by ensuring privacy, anonymity, and adherence to the platforms terms and conditions. Privacy and anonymity cannot be guaranteed due to potential de-anonymisation [29], however I will hold the minimum required data, whilst anonymising and aggregating data to mitigate this risk and ensure confidentiality, in-line with GDPR principles.

Beneficence and Non-maleficence
Additionally, although aiming to promote beneficence and uphold non-maleficence through improving pandemic forecasting, this can be undone without due-diligence. Misinterpretation and lack of reproducibility may lead to erroneous conclusions, whilst certain demographics may be underrepresented in training data, causing algorithmic bias and possibly leading to discriminatory predictions [3].

Transparency
Transparency about methodology, methods, data, and limitations is essential to mitigate unintended harm in real-world applications [3]. Transparency open identifies the constrains of my conclusions, whilst aiding reproducibility which in-turn supports the projects reliability [30].

Confidentiality
Measures will be taken such as anonymizing data and handling it securely. Transparency about methodology, methods, data, and limitations will be prioritized to mitigate unintended harm and ensure reproducibility. Overall, the research will uphold ethical standards outlined by the University of York

Due to ethical implications relating to social-media data, I have completed the University of York’s ethics fast-track application for authorisation before any data is acquired, and confirm it has been sent.

6 References

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

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[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] H. A. Park, H. Jung, J. On, S. K. Park, and H. Kang, "Digital Epidemiology: Use of Digital Data Collected for Non-epidemiological Purposes in Epidemiological Studies," *Healthc Inform Res,* vol. 24, no. 4, pp. 253-262, Oct 2018, doi: 10.4258/hir.2018.24.4.253.

[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] X. Hu *et al.*, "Event detection in online social network: Methodologies, state-of-art, and evolution," *Computer Science Review,* vol. 46, 2022, doi: 10.1016/j.cosrev.2022.100500.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] T. Szmuda, S. Ali, T. V. Hetzger, P. Rosvall, and P. Sloniewski, "Are online searches for the novel coronavirus (COVID-19) related to media or epidemiology? A cross-sectional study," *Int J Infect Dis,* vol. 97, pp. 386-390, Aug 2020, doi: 10.1016/j.ijid.2020.06.028.

[16] D. Lazer, R. Kennedy, G. King, and A. Vespignani, "The Parable of Google Flu:

Traps in Big Data Analysis," *Science,* vol. 343, 2014.

[17] Z. Tufekci, "Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls," in *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, ICWSM, Ed., 2014, vol. 14.

[18] 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.

[19] 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.

[20] 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.

[21] D. Wang *et al.*, "Development of an early alert model for pandemic situations in Germany," *Research Square,* 2023, doi: 10.21203/rs.3.rs-3108281/v1.

[22] 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.

[23] 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.

[24] 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.

[25] A. Sarker, S. Lakamana, W. Hogg-Bremer, A. Xie, M. A. Al-Garadi, and Y. C. Yang, "Self-reported COVID-19 symptoms on Twitter: an analysis and a research resource," *J Am Med Inform Assoc,* vol. 27, no. 8, pp. 1310-1315, Aug 1 2020, doi: 10.1093/jamia/ocaa116.

[26] 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.

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

[28] C. Klingler, D. S. Silva, C. Schuermann, A. A. Reis, A. Saxena, and D. Strech, "Ethical issues in public health surveillance: a systematic qualitative review," *BMC Public Health,* vol. 17, no. 1, p. 295, Apr 4 2017, doi: 10.1186/s12889-017-4200-4.

[29] L. Rocher, J. M. Hendrickx, and Y. A. de Montjoye, "Estimating the success of re-identifications in incomplete datasets using generative models," *Nat Commun,* vol. 10, no. 1, p. 3069, Jul 23 2019, doi: 10.1038/s41467-019-10933-3.

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