

Fast-Track Ethical Approval Form

This fast-track form is for **taught students** only. The form needs to be filled by the student and their advisor (or module leader if the project is a module assignment).

Research students and staff must complete the Full Ethical Approval Form.

Students: You need to discuss the ethical considerations of your project with your project advisor (or module leader if the project is a module assignment) and, if necessary, fill in a full ethics form to be submitted to the Physical Sciences Ethics Committee.

Advisors (or Module Leaders): You need to review and approve the completed form. Please ensure you are familiar with the University's [Code of practice and principles for good ethical governance](#) to guide your student effectively. Please seek guidance from the Departmental Ethics Officer(s) if you are uncertain about any ethical issue arising from this application.

Section 1. Potential Ethical Issues

Does your project involve any of the following? Please mark Yes or No for **all** issues.

No.	Issue	Yes	No
1.1	Human participants (adults or children)		X
1.2	Human data (e.g. data collected through surveys and questionnaires on issues such as lifestyle, housing and working environments, and attitudes and preferences, and datasets including human data)	X	
1.3	Applications that could potentially involve unethical practice, including potential dual-use applications (e.g. projects involving tools or data that can be used to attack systems)		X
1.4	Funding sources or collaboration with potential to adversely affect existing relationships or bring the University or Department into disrepute (e.g. projects related to gambling, dark market, etc.)		X
1.5	Restrictions on dissemination (e.g. not being allowed to publish certain datasets or results)		X
1.6	Military or defence context		X
1.7	Overseas countries under regimes with poor human rights record or identified as dangerous by the Foreign & Commonwealth Office		X
1.8	Human material (e.g. tissue or fluid samples), vertebrates, especially mammals and birds, or any other organisms not previously mentioned		X

If you answered **No** to all the above, you do not need ethical approval.

If you answered **Yes** to any of the above, you must complete this Fast-Track Ethical Approval Form, get it signed off by your project advisor (or module leader), and submit it for approval to the Departmental Ethics Officer(s).

This form is designed to verify specific conditions. If certain conditions are not satisfied, the form will guide you to complete a Full Ethical Approval Application, to be approved by the Physical Science Ethics Committee.

Section 2. Project Information

Student Name	Nathanael Bashford
Course Title	Masters in Computer Science and Data Analytics (Online)
Project Advisor	
Project Title	Investigating the Covid-19 Forecasting Capability of Covid-related Reddit posts using NLP and BERT.

Project Type:

<input checked="" type="checkbox"/>	Undergraduate Project	<input checked="" type="checkbox"/>	Postgraduate Project
<input type="checkbox"/>	Undergraduate Module Assignment	<input type="checkbox"/>	Postgraduate Module Assignment
<input type="checkbox"/>	Other - Please specify:		

Project Description

2.1. Provide a clear statement of your research questions or your experimental hypotheses.

Main 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?

2.2. Briefly explain what you are going to do in your study. Give sufficient detail that a non-expert in the subject can understand what you are proposing to do.

Brief Description:

I aim to explore the potential of using data from Reddit, a popular micro-blogging site, to improve forecasting models for Covid-19. The main goal is to see if classifying Covid-19 related posts on Reddit can help to improve disease forecasting projections beyond traditional data like 'daily infected cases' alone.

Data Collection:

I will retrieve my data from a pre-collected dataset collected by the 'SocialGrep'. This dataset is freely available on the website www.socialgrep.com, and the use of this data available for academic use as stated under the license: CC BY 4.0 (References below for website containing the dataset, and the License)

Website: SocialGrep (2021) *The Reddit COVID Dataset*. Available at: <https://socialgrep.com/datasets/the-reddit-covid-dataset> (Accessed 26 November 2023). (The TASL Method of Attribution for citing Creative Commons Licensed work:)

License: "The Reddit Covid Dataset." SocialGrep. <https://socialgrep.com/datasets/the-reddit-covid-dataset>. Licensed under a Creative Commons Attribution 4.0 License (<https://creativecommons.org/licenses/by/4.0/>)

This dataset contains all Reddit posts/comments that state the term 'Covid' up until October 25, 2021.

I will additionally gather historical data on Covid-19 cases in the USA for the same period, freely available from the World Health Organisations (WHO) website.

Data Analysis:

Reddit data will be cleaned before Natural Language Processing (NLP) techniques will be used to classify Reddit Posts/comments. Data cleaning will include cleaning text (removing irrelevant information - URLs, emoji's etc.) then classifying the cleaned content into different categories based on the context, like whether a user is talking about having

Covid-19 or discussing someone they know who has it.

To classify these discussions, I will use a method known as BERT (Bidirectional Encoder Representations from Transformers), an advanced NLP technique, ideal for accurately understanding context of texts.

Forecasting Model Development:

I will develop forecasting models called LSTM (Long Short-Term Memory) models. These models can understand complex patterns in data over time and can make predictions of future trends.

The models will be trained using both the Reddit data and traditional Covid-19 incidence (cases) and determine if the combining of these two sources improves the accuracy of the forecasts.

Comparative Analysis:

The accuracy of the forecasts from models using only traditional Covid-19 data will be compared against models that additionally include the Reddit derived data. This will help understand if Reddit data adds value to the forecasting.

Evaluation:

The effectiveness of these models will be assessed using statistical measurements like Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE)

Purpose:

The study aims to explore the effectiveness of integrating social media data into disease forecasting models. Platforms like Reddit have become rich sources of real-time public sentiment and behaviour. Understanding if this data can enhance our ability to forecast diseases like Covid-19 could be crucial in preparing and responding to public health emergencies more effectively.

Section 3. Informed Consent

If you do not have participants in your study, please mark the following and skip to Section 4.

X	There are no participants in my study.
---	--

If you have participants in your study, complete the following table.

If you answer **No** to any of the following, you must submit a Full Ethical Approval Form.

No.	Question	Yes	No	N/A
-----	----------	-----	----	-----

3.1	Will you inform the participants of the purpose of the study, the investigators, and any funding source?			
3.2	Will you describe the procedures and the data collected to participants in advance, so that they are informed about what to expect? Note: any personal or sensitive information, and any audio or video recording must be explained explicitly and in detail.			
3.3	Will you tell participants that their data will be treated with full confidentiality and that, if published, it will not be identifiable as theirs? Note: If there is a plan to publish data, e.g. direct quotations, on an individual basis (rather than in aggregate), the exact anonymisation method needs to be explained clearly.			
3.4	Will you inform participants of any possible risks of participation?			
3.5	Will you inform participants that their participation is voluntary, that they may withdraw from the research at any time (even after the data collection, for a reasonable and specified time period) and for any reason without any negative consequences, and how they can request withdrawal?			
3.6	Will you obtain and record explicit consent for participation?			
3.7	If the research is observational, will you obtain and record explicit consent to being observed?			
3.8	If the research involves audio or video recording, will you obtain and record explicit consent to being recorded?			
3.9	If there is a plan to publish data, e.g. direct quotations, on an individual basis, will you obtain and record explicit consent to such publication?			
3.10	If you wish to contact participants in the future, will you obtain and record explicit consent to being contacted with the contact method specified?			

If you answered **Yes** to any of the above, this must be explicit in your supporting documents, e.g. consent forms, information sheets, and questionnaires. You need to submit these supporting documents along with this form.

Section 4. Protocol Issues

If you answer **Yes** to any of the following, you must submit a Full Ethical Approval Form.

No.	Question	Yes	No	N/A
4.1	Is your study designed to be challenging or disturbing (physically or psychologically) for anyone including yourself?		X	
4.2	Will you deliberately mislead your participants?		X	
4.3	Does your study involve taking bodily samples?		X	
4.4	Is your study physically or psychologically invasive?		X	
4.5	Is there any obvious or inevitable adaptation of your research findings to ethically questionable aims?		X	
4.6	Could the methodologies or findings of your study damage the reputation of the University of York?		X	

Section 5. Health and Safety

Please identify any risks to the participants and state any precautions you will take to ensure their physical and mental health and safety. If you believe there are no risks, please state why.

There are no health and safety risks as there are no participants in this project. Reddit data collected from the pre-collected dataset is already anonymised and contains only minimal information. Data will additionally be aggregated into the number of Reddit posts/discussions that fall into each classification category (i.e. Daily posts stating someone has Covid, Daily posts stating someone the author knows has Covid) in this way, the minimum required amount of data will be stored, and only aggregated data will appear in the final project.

Section 6. Vulnerable Groups and Animals

If you answer **Yes** to any of the following, you must submit a Full Ethical Approval Form.

If your participants are **patients**, in addition to the Full Ethical Application you must follow the Guidelines for Ethical Approval of NHS Projects.

If your project involves any vulnerable group, you may also need to obtain satisfactory Disclosure and Barring Service (DBS) clearance (or equivalent for overseas students).

No.	Question	Yes	No	N/A
6.1	Does your project involve working with animals?			X

6.2	Does your project involve any of the following vulnerable groups?			X
	<ul style="list-style-type: none"> Children under 18 			X
	<ul style="list-style-type: none"> People with learning difficulties 			X
	<ul style="list-style-type: none"> People who are unconscious or severely ill 			X
	<ul style="list-style-type: none"> Patients, e.g. NHS patients 			X
	<ul style="list-style-type: none"> Other vulnerable groups – specify below <div style="border: 1px solid black; height: 50px; width: 450px; margin-top: 10px;"></div>			X

Section 7. Data Protection

If you answer **No** to any of the following, you must submit a Full Ethical Approval Form.

No.	Question	Yes	No	N/A
7.1	Any personal or sensitive data will be stored in password protected folders on computers controlled by the University or an approved partner.	X		
7.2	Any hard copies of personal data (including any hard-copy consent forms) will be stored in a secure place within University premises, or those of an approved partner.	X		
7.3	Only the student and advisor will have access to the data generated from the study. Note: The advisor may share the anonymised data with other researchers at the University of York, but the consent form needs to make this clear.	X		
7.4	The data will be preserved beyond the study in line with University policy and will be placed in the custody of the advisor at the end of the project, or destroyed in accordance with information provided to participants.	X		
7.5	All data will be anonymised prior to analysis. Please state your method of anonymisation: <div style="border: 1px solid black; padding: 5px; margin-top: 10px;"> The prior collected dataset is already anonymised, containing only the following data: Type (Reddit 'Comment' or 'Post'), id </div>	X		

	<p>(unique base-36 id for each comment/post), subreddit.name (Human-readable name of the comment's subreddit), created_utc (Timestamp of the comments creation), permalink (link to the comment on Reddit), body (The comments/Posts text), sentiment (analysed sentiment of the content), Score (the comment's score).</p> <p>The only data columns that will be kept after collection will be: Type, id, subreddit.name, timestamp, body.</p> <p>None of the columns available contain identifiable information, and only necessary text from the body of each comment/post will remain after data cleaning. After cleaning, as mentioned previously,</p> <p>As stated above, data will be aggregated into the number of Reddit posts/discussions that fall into each classification category (i.e. Daily posts stating someone has Covid, Daily posts stating someone the author knows has Covid) in this way, the minimum required amount of data will be stored, and only aggregated fully anonymised data will be analysed and appearing in the final project.</p>			
--	--	--	--	--

Section 8. Further Project Information

If you have participants in your project or if the data you are working with is about humans, briefly specify each of the following.

8.1. Who are the target participants or data subjects? And how are you recruiting them or how were the data about them collected?

The focus in this project is on the textual content of posts and comments on Reddit that mention Covid-19, rather than on individual users/subjects themselves.

There is no recruiting of any participants in this study, but the Reddit data I aim to use has already been collected and aggregated by 'SocialGrep'. 'SocialGrep' is a website that previously provided users free term/query searches from historical Reddit posts, but now charges users for historical access. It does however provide access to several pre-collected datasets, and the dataset I aim to collect from their website is titled: 'The Reddit COVID Dataset', and contains all posts and comments found to mention the term 'COVID' prior to October 25, 2021. It is in the form of two CSV files, the first contains 4.5 million 'Posts' that contain the term 'COVID', and the other file contains 17.7 million 'Comments' that contain the term 'COVID'.

- from the website 'SocialGrep.com' which collected/generated the dataset, it states in the description the Selection criteria and the Date range for the selected Posts/comments

(Reference: Website: SocialGrep (2021) *The Reddit COVID Dataset*. Available at: <https://socialgrep.com/datasets/the-reddit-covid-dataset> (Accessed 26 November 2023)).

The freely available dataset has the following specified licensing regarding the use of this dataset (License): Attribution 4.0 International (CC by 4.0).

The reference for this license in TSLA format is: "The Reddit Covid Dataset." SocialGrep. <https://socialgrep.com/datasets/the-reddit-covid-dataset>. Licensed under a Creative Commons Attribution 4.0 License.

Below states what this license allows and is copied from the Creative Commons (CC) website:

CC BY 4.0 DEED

Attribution 4.0 International

You are free to:

Share — copy and redistribute the material in any medium or format for any purpose, even commercially.

Adapt — remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

Attribution - You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

No additional restrictions - You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable exception or limitation.

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as publicity, privacy, or moral rights may limit how you use the material.

The above text and the full License (Deed) is available from the following website: (<https://creativecommons.org/licenses/by/4.0/>)

8.2. Exactly what pieces of personal and demographic information you are collecting or using from an existing dataset? Explain why you need each.

As previously stated, the prior collected dataset contains only the following data: Type (Reddit 'Comment' or 'Post'), id (unique base-36 id for each comment/post), subreddit.name (Human-readable name of the comment's subreddit), created_utc (Timestamp of the comments creation), permalink (link to the comment on Reddit), body (The comments/Posts text), sentiment (analysed sentiment of the content), Score (the comment's score).

The only data columns that will be kept after collection will be: Type, id, subreddit.name, timestamp, body.

None of these columns contain personal and demographic information, except for what may be mentioned within the body ('Text') of the total 22 million combined posts/comments.

For my project I need to classify each post/comment into whether the text states or implies that an individual (author) has/has Covid, or whether someone the author knows has/has Covid. This will be generated into time-series data by aggregating how many posts/comments fall into each of these categories each day.

Demographic information therefore is not of interest - if it is contained at all within the text – and will not be analysed.

Personal information relating to an individual having Covid is of interest as this will be used to classify text into one of the categories I wish to classify text. This personal information is already anonymised, and every effort will be taken to ensure that no identifiable information is present or stored, and again identifiable information is of no interest to this study.

As stated above, data will be aggregated into the number of Reddit posts/discussions that fall into each classification category (i.e. Daily posts stating someone has Covid, Daily posts stating someone the author knows has Covid) in this way, the minimum required amount of data will be stored, and only aggregated fully anonymised data will be analysed and appearing in the final project.

8.3. How are you planning to publish your data and analysis results: on an aggregate basis or on an individual basis? If on an individual basis, explain why it is **necessary** to do so and if they will be anonymised.

I am publishing my data in an aggregated way, as the focus of my project is on the predictive ability of the classified Reddit Posts/comments, and this is aggregated into time-series data, stating the daily total of 'Texts' that belong to each classification category.

I will therefore publish and analyse the spread and distribution of this aggregated time-series data, but it is always aggregated and never at an individual basis.

8.4. Is there a risk that the data or analysis results you publish can be de-anonymised? If yes, explain how you are planning to mitigate it.

No personal or identifiable information will be published. Only aggregated 'tallies' of how many Posts/Comments ('texts') were classified into each class by day/week etc.
No text from the retrieved Reddit dataset will be published, only mentions of the text cleaning that will occur will be published (for reproducibility), but this will refer to the cleaning of all texts, rather than referring and publishing examples of specific texts in this project.
For example, a table/graph might be published showing the number of texts classified as 'User has Covid' correlated or plotted against real covid incidence, to highlight the relationship between the two, revealing no information that may be de-anonymised.


8.5. If you have questionnaires or interviews, will you give participants the option of skipping any questions they do not want to answer? If any of your questions is/are mandatory, explain why it/they must be.

N/A

Section 9. Student Declaration

Please mark boxes below that apply and sign.

X	I have considered the ethical implications of this project and have identified no significant ethical implications requiring a full ethical application submission to the Physical Sciences Ethics Committee.
	I will include all supporting documents (e.g. consent form, information sheet, questionnaires, interview schedules) with this application.
	All the components specified in Section 3 are included explicitly in the information sheet / consent form submitted with this application.

Student Name	Nathanael Bashford
Student Signature	
Date	15 th November 2023

Section 10. Advisor (or Module leader) Approval

If this project is related to any previously approved work, please include the reference number(s) for that work here.

--

Please mark boxes below that apply and sign.

X	The student has taken all reasonable steps to ensure ethical practice in this study and I can identify no significant ethical implications requiring a full ethical application submission to the Physical Sciences Ethics Committee.
	I have checked and approved all supporting documents required for this application.
	I understand that completion of this form indicates that from the ethical point of view I am willing to share responsibility for the work being conducted.

Advisor Name	Waseem Ahmad
Advisor Signature	Waseem Ahmad
Date	21/11/2023

Submission

Please make sure this form is completed in full, signed by both the student and the advisor, and accompanied by required supporting documents (e.g. participant information sheet and consent forms). Otherwise, the approval process is delayed.

Please submit this form and all supporting documents, preferably all in pdf unless other file types are appropriate, to the Departmental Ethics Officers through the submission form listed in the [Ethics in Student Projects](#) information page.