Literature Exploration Analysis Platform (LEAP) Master's Project Terms of Reference

DRAFT

Jon-Paul Boyd
School of Computer Science and Informatics
De Montfort University
United Kingdom

I. Summary

The volume of COVID-19 research literature is doubling every 20 days. Such rapid growth makes it outright impossible for humans to manually evaluate all available information. This project aims to architect, develop and evaluate a tool in which subfields of artificial intelligence, including natural language processing (NLP) and deep learning (DL), are leveraged to search, mine and explore COVID-19 related unstructured academic literature and social media data. Features including Q&A based knowledge discovery, automated literature summarization, topic classification and clustering, entity graph analysis and time series visualisation will provide decision-making support to medical researchers so the best-informed next stage research actions can be taken.

II. Background

"*Coronaviruses are a large family of viruses which may cause illness in animals or humans. In humans, several coronaviruses are known to cause respiratory infections ranging from the common cold to more severe diseases such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). The most recently discovered coronavirus causes coronavirus disease COVID-19.*" [1].

"*This new virus and disease were unknown before the outbreak began in Wuhan, China, in December 2019. COVID-19 is now a pandemic affecting many countries globally.*" [1]. "*The common symptoms of COVID-19 are fever, dry cough, and tiredness. Less common symptoms can include aches and pains, nasal congestion, headache, conjunctivitis, sore throat, diarrhoea, loss of taste or smell or a rash on skin or discoloration of fingers or toes.*" [1]. "*Most people (about 80%) recover from the disease without needing hospital treatment. Around 1 out of every 5 people who gets COVID-19 becomes seriously ill and develops difficulty breathing. Older people, and those with underlying medical problems like high blood pressure, heart and lung problems, diabetes, or cancer, are at higher risk of developing serious illness. However, anyone can catch COVID-19 and become seriously ill.*" [1]. As of 26th June 2020, globally there have been 9, 611, 062 confirmed cases of COVD-19 resulting in 489, 343 deaths [2].

In the first week of May alone more than 4000 new papers were published that were relevant to a virologist studying COVID-19 [3]. Literature publication is doubling every 20 days, among the biggest explosions of scientific literature ever [3]. Researchers do not have enough time to read and evaluate all available articles. The rush to publication has impacted the quality of literature, with many researchers posting preprints without peer review, content with commentary only, poor quality modelling or no original findings.

On March 16th, 2020 the *COVID-19 Open Research Dataset (CORD-19)* of machine-readable scholarly literature about COVID-19, SARS-CoV-2, and the Coronavirus group was made publicly available [4]. It contains containing over 167,000 scholarly articles, with 79,000 of those full text. Papers are both historical, dating back to the 1950's, and cutting-edge contemporary research. The dataset was compiled by a partnership between the Allen Institute for AI, the Chane Zuckerberg Initiative (CZI), Georgetown University's Centre for Security and Emerging Technology (CSET), Microsoft, the U.S. National Library of Medicine (NLM) and U.S. National Institutes of Health. The purpose in releasing the dataset was to provide a collection of relevant COVID literature on which the global research community could apply natural language processing and other A.I. techniques to generate new insights in support of the ongoing fight against this infectious disease. There is a growing urgency for these approaches because of the rapid acceleration in new coronavirus literature, making it difficult for the medical research community to keep up.

III. PROPOSED WORK

A. *Aims and Objectives*

The main aim of this research project is the development of an AI-based academic literature search engine with question-answering, analytic, and exploration features. It will put natural language processing, data mining, machine learning and other approaches to use in supporting the medical community answering questions about the pandemic and connecting insights across research included in the CORD-19 dataset and other related data sources.

B. *Sub Objectives*

Note the organisation of sub objectives as hypothesis (H*n*) grouped by research question (RQ*n*).

RQ1 Feasibility - Can such an unstructured data search and analytics application be delivered through to execution?

- **H1** Discover the available open source modules that can expedite such an application build
- **H2** Determine cloud hosting providers and services that support data store, Python FaaS (Functions as a Service) and NLP
- **H3** Estimate cloud hosting costs
- **H4** Consider rapid prototyping of main features

RQ2 Functionality - Can A.I. enable intelligent search, summarisation, and data-mining analytics of unstructured literature?

- **H5** Verify NLP sensibly transforms natural language search queries such that only the most relevant literature is listed
- **H6** Confirm A.I. algorithms can rank listed literature according to their search relevancy
- **H7** Ascertain A.I. algorithms can generate a relevant brief summary per document from a medical researcher perspective
- **H8** Can A.I. extract a high quality "one page" summary from all literature relevant to a search query?
- **H9** Can data mining be used to develop insights, findings, patterns, emerging trends and tables of facets from the literature?
- **H10** Determine if graph analysis can build a network of papers, drugs, authors, affiliations, sources and trends
- **H11** Can A.I. facilitate associative exploration by making new suggestions and propose other papers based on initial search

RQ3 User Experience - Can the application effectively support COVID-19 research?

- **H12** Verify application delivers research value and enables better next-step research decisions
- **H13** Ascertain if the application reduces the overall manual time spent on COVID-19 research
- **H14** Check if it is possible to visualize publications over time
- **H15** Determine if the application enables researchers to react to emerging trends faster
- **H16** Verify if query, search results, summaries, data mining visualization, trending social media and news feeds can be presented in a one-stop researcher cockpit

RQ4 Technical Performance

- **H17** Verify query search and result presentation completes in acceptable timeframes using full dataset
- **H18** Verify all interactive elements including time series adjustment, network graph node and facet selection perform

IV. METHODOLOGY

The project process will begin with planning ways of working with my supervisor – establishing methods on how and when we will communicate to assess progress and discuss open points. This ToR will be discussed, revised and "*signed-off*" to provide an agreed framework and delivery scope. A comprehensive literature review will follow, to include the background and latest on the COVID-19 pandemic itself, the state of COVID-19 research including from academic literature and other sources, and current literature research tools. Research the particular semantics of medical research query. The review will additionally include an evaluation of techniques, tools, algorithms, and models used for unstructured text search and information extraction. The existing body of knowledge will be researched for relevant papers.

A mixed-methods, after-only study design will be employed to evaluate the research. Qualitative in-depth user feedback pre and post experiment from questionnaires containing closed (e.g. "*Are you a medical researcher?*" – dichotomous variable with yes/no) and open questions ("*What is your impression of the search results and researcher cockpit?*") will enrich quantitative empirical measurements collected by the technical framework. The technical solution itself will constitute the main research instrument. Observations will be made on each of the 4 concepts outlined in "*sub objectives*". RQ1 Feasibility will be measured with scales qualitative in nature, as will RQ3 User Experience (except H13 measured in time interval scale). RQ2 Functionality is mixed: H5 can be quantitively measured by ordinal scale (e.g. agree, neutral, disagree) whereas H9 is qualitative. RQ4 Technical Performance is quantitative.

V. PROGRAMME OF WORK – WORK PACKAGES WITH DELIVERABLES

An interactive question-answering and data mining tool will be developed to enable the medical community to explore COVID-19 academic literature and other relevant data sources. Its purpose is to help them find the most relevant information they need quickly, identify patterns and gain insight from it by way of a user-friendly interface comprising ranked lists, tagged documents, graphical data visualisation and interactive exploration, all in a one-page research cockpit. The tool will be architected as a cloud-hosted application on the Microsoft Azure platform. The identified objectives will be delivered by the following work packages (WP*n*).

WP1 Literature Review (tbd weeks)
Examine the current body of knowledge as elaborated in "*Methodology*".

WP2 Accounts (1 week)
Procure appropriate access accounts for cloud deployment and data scraping.

WP3 Feasibility Study (tbd weeks)
For such an ambitious project, the project will rely heavily on open source data, backend modules, algorithms, and frontend frameworks, all of which should be identified and evaluated upfront. Furthermore, appropriate cloud providers need to be identified and deployment costs estimated. Certain concepts may need to be proven feasible with small prototype test functions.

WP4 Prototype Core Data Ingestion (tbd weeks)
Create a digital repository of academic data by uploading the CORD-19 dataset. Build prototype tooling, primarily as callable functions, that facilitate the ingestion, pre-processing, filtering, indexing, tagging, classification, and metadata creation of the main CORD-19 dataset. Filtering is especially important to exclude poor quality sources (commentary only with missing main body, irrelevant documents that do not include essential key terms such as coronavirus) from being ingested into the database.

WP5 Prototype Search (tbd weeks)
Build prototype search tool, initially using an identified data subset to simplify and expedite evaluation. Develop functions for query parsing, result filtering and result ranking.

WP6 Prototype Results (tbd weeks)
Build prototype functions to manage, aggregate and analyse query results, including summarisation, classification, and clustering. The results will be ranked relevant to the asked question. Each result item will be tagged with several keywords that include source (academic journal name or news channel for example), topic labels, category, year published, peer-review status etc.

WP7 Prototype Supplementary Data (tbd weeks)
Build plug-ins to connect and scrape 3rd party sources including Twitter, news channels, ResearchGate, Google Scholar and others. Such scraping of web-based sources supports ranking mechanisms (citations, year last published, public praise and critique of publications), and provides additional information for the data mining search tools, metrics, and trending panels. This is intended to bring silos of COVID information together into one collection which is vital for best results. Implement a function to allow upload of additional literature not part of CORD-19 dataset.

WP8 Web UI (tbd weeks)
Using mock-up tools design the research cockpit frontend. Build, deploy and test the web UI.

A search box will facilitate the asking of a question in human natural language (English), for example "*What is known about transmission, incubation, and environmental stability?*". A short snippet of a result item will be included, highlighting the terms/key passage of relevance.

Filters for source (journal, social media platform, news channel etc), author, category, key terms etc will facilitate iterated refinement of results. Facet lists will aggregate specific search result themes, including topic frequency (epidemiology, vaccination, antibody, paediatrics), affiliation (De Montfort University, Oxford University, Johns Hopkins etc), which can be selected to refine the search results further, effectively enhancing the original query with additional search concepts.

A timeline will show the number of articles published over time according to the selected facets (topic frequency, affiliation, author etc). This will give an indication of emerging trends and directions. Adjusting the timeline (from/to date) will update the article inclusion, and hence, the facet lists.

An interactive clustering chart will allow analysis of similarity and distance between literature. A topic chart will show a word bubble with common words across all result documents sized by frequency. Similar topic chart word bubbles will be available for a selected result item, to summarize the complete document or abstract only.

A world map will show tagged countries according to original query results and interactive filtering.

UI evaluation will be supported by crowd-sourced medical and UX experts.

WP9 Prototype Supplementary Features (tbd weeks)
Implement search autocomplete and autocorrect, result download and email in Excel format. Integrate social media paper peer review and author citation count into ranking algorithm. Implement paper sentiment analysis.

WP10 Industrialise (tbd weeks)
Code refactor and clean-up of prototyped functions and other components, conforming to own and industry standards. Implement application security including application log-in.

WP11 User, Technical Evaluation, Test Results (tbd weeks)
Researcher end user evaluation of the solution, including interviews and observations. Includes writing on-line end user manual. Data collection to measure each hypothesis including technical performance testing.

WP12 Technical Design Document (tbd weeks)
Detailed documenting of the technical implementation, including architecture blueprint.

WP13 Demonstration (1 week)
Creation of video demonstration

VI. ACADEMIC AIMS

- Understand the challenges of unstructured data search and analytics, with an initial focus on medical domain semantics
- How to create, aggregate and search a digital repository of unstructured data
- Implementation of filtering mechanisms to exclude documents according to some criteria
- Text pre-processing to exclude stop words (and, is, are, at, has etc) and punctuation (apart from social media channels where smiley or sad face indicates sentiment)
- Text pre-processing to implement stemming to reduce words to their base stem (e.g. diseas is stem for disease and diseases)
- Learn how to build plug-in connectors for web-scraping of additional unstructured data sources including social media platforms, research platforms and news channels
- Learn how to curate the digital information store by additional additional properties including a peer-review check (mitatge abstract oversell etc), author citation count
- Learn how text documents including academic literature, tweets and news feeds can be digitally stored, indexed (BM25), classified (FastText), tagged, ranked and filtered.
- Learn how models can be trained on unstructured data (as opposed to continuous, numerical data) to understand the human language.
- Learn how encoder-decoder, causal, self-attention and transformer mechanisims can be used to summarize text and build Q&A systems.
- Learn how adidtional metadata per document can be extracted, stored and used, to support summarization and classification for example
- Learn how hidden Markov models and word embeddings can implement autocorrect and autocomplete when entering search query
- Learn how state-of-the-art T5 and BERT models that learn contextual relations between words can better support Q-A in the medical search domain
- Learn how mechanisms such as logistic regression, neural networks, LSTMs (Long Short Term Memory) networks, GRUs (Gated Recurrent Unit) networks ands siamese networks can be used for sentiment analysis and named entity recognition (categorization)
- Learn how to cluster unstructred text by contextual pattern
- Learn how to design, build, and deploy a user friendly interface for researchers
- Learn about distributed, resource scaling, micro-services based application development on the Azure cloud platform
- Ideally the application can be designed in such a way that it can be useful beyond the COVID-19 pandemic and prove useful for unstructured search in other domains.

VII. RESEARCH QUESTIONS

The application will enable medical researchers to find quickly find answers to important questions about COVID-19, aggregating, summarising and ranking unstructred data from the large CORD-19 dataset which will be supplemented with complimentary data from other relevant channels. It should enable researchers to efficiently focus their efforts on the highest quality papers most relevant to them.

The Kaggle COVID-19 Open Research Dataset Challenge [4] proposes a list of key scientific questions compiled by the NASEM's SCIED (National Academies of Sciences, Engineering, and Medicine's Standing Committee on Emerging Infectious Diseases and 21st Century Health Threats) research topics and the World Health Organization's R&D Blueprint for COVID-19. The application should be able to answer the proposed list, which includes the following examples:

- Summary of articles that address relevant factors related to COVID-19. Specifically, what the literature reports about hypertension, diabetes, male gender, heart disease etc.
- Summary of articles that address therapeutics, interventions, and clinical studies. Specifically, the effectiveness of drugs being developed and tried to treat COVID-19 patients, Methods evaluating potential complication of Antibody-Dependent Enhancement (ADE) in vaccine recipients etc.
- Summary of articles that address diagnostics for COVID-19. Specifically, what do we know about diagnostics and coronavirus, Development of a point-of-care test and rapid bed-side tests etc
- What is known about transmission, incubation, and environmental stability?
- What do we know about COVID-19 risk factors?
- What do we know about vaccines and therapeutics?
- What do we know about virus genetics, origin and evolution?
- What has been published about medical care?
- What do we know about diagnostics and surveillance?

VIII. RESEARCH LITERATURE

The following topics will be covered in the literature review:

- The COVID-19 disease itself and it's impact on humanity.
- The state of COVID-19 academic and other literature, including timelines, growth, availability, quality, direction, trends and it's consumption by varying users and tools.
- The challenges facing epidemiologists, virologists and other researchers searching, consuming and evaluating COVID-19 literature.
- The benfits of an automated, AI-based search and exploration tool to faciltate COVID-19 research
- The specific semantic nature of a medical human natural language query (word order, meaning etc)
- Natural Language Processing (NLP) – background, common and state-of-art approaches. Common practices, steps and flow (pre-processing, stemming, indexing, tagging, search, auto-complete, auto-correct, ranking, classification)
- Machine learning for state-of-art NLP, including BERT for document context, transfer learning
- Cloud computing – benefits, architecture, applicable services, performance, scale, reach, security, cost

IX. JUSTFICATION OF RESOURCES

People The research author alone will be responsible for the study design and execution., including all aspects of the research instrument. User experience (UX) testing will be crowd sourced. Regular review sessions (frequency to be agreed) with project supervisor.

Data The primary CORD-19 dataset of academic, unstructured text literature is freely available. Secondary data including social media feeds, research platform feeds (ResearchGate, Google Scholar) and news channels will only be used where terms and conditions of use are fully satisfied.

Hardware Microsoft Azure cloud hosted, normally on paid subscription / hourly rate.

Software Open source freely available, including Python and appropriate modules.

Technical references Many Python books in personal library. Additional free learning materials available online.

Development Languages – Primarily Python as by far most popular, well-supported language for A.I. and M.L.

Front-end technology tbd.

Development Tools – PyCharm Professional (have student license)

X. RISK ASSESSMENT

The following have been identified as risks:

- **Cloud computing costs** – Consuming storage and computing resources on Azure will incur costs. Will mitigate by developing and testing as much possibly locally before cloud deployment. Will investigate possible Azure Cloud sponsor to help with costs.

- **Growing number of COVID-19 literature behind paywalls** – Recent study concluded that a growing number of COVID-19 literature studies are behind paywalls. Therefore the solution will rely on the primary data being sourced from the freely available CORD-19 dataset

- **Terms of service for web scraping and using 3rd party data** – Accounts used to scrape data from social media and research platforms may violate terms of service, need to clarify.

One critical goal is to deliver a user experience that fully engages. This ideally will require the participation of real medical researchers. Possible mitigation in finding these expert users can be by crowd-sourcing testing via social media. The main financial cost will be cloud hosting of the solution throughout development. This will be funded personally although some form of sponsorship would be wlecome. I have mitigated any data privacy concerns by opting to uyse datasets in the public domain.

XI. REFERENCES

[1] W. H. Organisation, “Q&A on coronaviruses (COVID-19),” [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/q-a-coronaviruses.

[2] J. H. University, “COVID-19 Map,” [Online]. Available: https://coronavirus.jhu.edu/map.html.

[3] Sciencemag.org, “Scientists are drowning in COVID-19 papers. Can new tools keep them afloat?,” [Online]. Available: https://www.sciencemag.org/news/2020/05/scientists-are-drowning-covid-19-papers-can-new-tools-keep-them-afloat.

[4] Kaggle, “COVID-19 Open Research Dataset Challenge (CORD-19),” [Online]. Available: https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge.

[5] C. Stack.AI, “GitHub - Cortical Stack - heuristic optimization platform repository,” [Online]. Available: https://github.com/corticalstack/heuristic-optimization-platform.