Literature Review

Applying Machine Learning to improve the prediction of future COVID-19 cases in Ireland

Meet Mehta

Patrick McLoughlin

Simon McLain

February 13, 2022

Table of Contents

**Contents**

[**1.** **Introduction** 4](#_Toc94722857)

[**1.1.** **<<** **Core technical area 1>** 4](#_Toc94722858)

[**1.1.1.** **Sub technology 1 - core technical area 1** 4](#_Toc94722859)

[**1.1.2.** **Sub technology 2 - core technical area 1** 4](#_Toc94722860)

[**1.2.** **Big Data** 4](#_Toc94722861)

[**1.2.1.** **Healthcare Big Data analytics** 5](#_Toc94722862)

[**1.2.2.** **Demographic and Socio-economic Big Data** 5](#_Toc94722863)

[**1.2.3.** **Social Media Big Data** 5](#_Toc94722864)

[**6.** **Conclusion** 5](#_Toc94722865)

[**Works Cited** 6](#_Toc94722866)

Table of Figures

# **Introduction**

<< This section should outline the context of the work. What is the real world problem you are trying to address? What are the quantifiable issues which need to be overcome?>>

<<The section should state the major categories within which the literature review will be scoped>>

<<The section should provide high level reference into which the subsections interrelate>>

# **<<** **Core technical area 1>**

<< For core technical area 1 this section should outline the relevance of the technical area to the overall research agenda. Why is this technical area being discussed? How can this area be sub-categorised? How can these technologies address the overall challenge discussed in section 1?>>

# **Sub technology 1 - core technical area 1**

# **Sub technology 2 - core technical area 1**

# **Big Data**

Big Data is defined as “data that contains greater variety, arriving in increasing volumes and with more velocity” [1]. In the past data was stored mostly on paper records of local digital storage devices, if at all. As such the availability of data was much reduced when compared with today. In essence the dramatic increase in data collection then storage in a shareable way as potentially provides scientists a wealth of insights. With this technological leap scientists can enrich the data they collect themselves through observations and experiments with open-source data available collected from a multitude of organisations and institutions.

Epidemiologists undertake systematic, data-informed, investigations into the biology of pathogens and their complex interactions with their hosts to identify changes in behaviour that could contain an outbreak reduce or even eliminate the spread of disease. Big Data can potentially be utilised to enrich their knowledge and help build reliable predicts of the development of an epidemic as well as the effectiveness of control measures.

Potentially valuable Big Data artifacts can be broadly placed in three categories.

Firstly, healthcare Big Data relevant to our project will include detailed information collected from patients who have presented with disease. It will provide information related to age, gender and may be enriched with data relating to medical health, pre-existing conditions, vaccination status etc. This information is of course collected after outbreaks of the disease have occurred, but it is specific to the disease as it will contain information for individuals who have been proven to be infected. The data is normally collected by health care providers and government agencies. This data is useful for identifying risk factors such as age and gender. As it is collected in real time during an outbreak situation the data will evolve over time, as more knowledge of the disease is gained.

Other healthcare datasets can also be of use, for instance, if certain underlying conditions (e.g., cardiovascular disease) are risk factors for diseases, such as COVID-19, becoming life threatening then knowledge of their prevalence in the community may also clarify the severity of the disease.

The next relevant Big Data category is s population demographics. Reliable demographic information will typically be collected by governments through a census. It describes the population geospatially, by ethnicity, religion, gender, and age. Of course, this data is not specific to the outbreak of a disease, however it can provide valuable contextual information. For instance, demographic data can help to understand population density, a factor that will be important for the spread of airborne diseases. In normal circumstances this data will not change significantly over a period of several months. So, the most recent census data can be taken as generally representative of the composition of the population.

The final relevant Big Data category is Social Media data. This data can come from various sources such as, Google Searches, posts to Twitter, Facebook, Instagram etc. This data is typically unstructured, taking the form of text, emoji, gifs, pictures, and documents. Social media data has the advantage of being available in real time and being representative of the cognizance and behaviours of the actors participating within the Social Media space. Relevant applications in the context of the spread of COVID-19 could be Google searches for testing centres, pharmacists, and symptoms of the disease. The behavioural indicators can include hashtags that indicate compliance or non-compliance with public health advice, as well as number of people contained within photographs. As such this data may be helpful in understanding whether or not a community is engaging in behaviours that would increase or inhibit the spread of disease.

# **Social Media Big Data**

[2] Use data from Google Trends and Twitter to detect influenza in Greece. Hardware included a computer with uninterrupted power supply. An ARIMA(X) and a custom prediction model were coded in the Python language. Their research demonstrated that internet data can be useful in tracking epidemics with the potential to estimate and predict influenza. Twitter REST API and the Twitter Streaming API can be used to gather tweets, but are rate limited reducing the amount of data that can be gathered. Google data can be accessed through their web application as well as using API’s such as PyTrends.

[3] propose improve predictions by using neural network based Long Short-Term Memory models and auto-regressive moving average models with two channels of inputs to incorporate social media and historic disease data. The keras2 python library is used to implement the machine learning algorithm. The approach concluded that social media data can be used in combination with current case counts to improve the accuracy of prediction and proposed enriching the dataset with addition information such as climate data.

[4] Gathered and pre-processed English language and Tagalog language Flu related tweets through the Twitter API. Pre-processing included removal of punctuations, special characters, hashtags. Programming was carried out using the Python programming language and associated libraries. Including the use of the Natural Language ToolKit [NLTK] for stemming and lemmatisation. Machine Learning algorithms were implemented using the Scikit Learn library. Their goal was to apply the C5.0 decision tree algorithm to improve on predictions made using the Naïve Baynes algorithm. Their implementation of the C5.0 model had a significant increase in accuracy when compared to Naïve Baynes, but was more complex in terms of Big O.

[5] Discuss the failure of researchers at Google to use Big Data to accurately predict flu outbreaks, particularly how the model was initially accurate but then began to perform worse over time. An example citing being the algorithms’ weakness to overfitting on seasonal factors unrelated to influenza, such as “high school basketball”. In their research the authors highlight both the benefits and challenges of using Big Data for the public good. They do this in terms of the selection and pre-processing of the data so that it contains components with a direct, real world, association with the dependent variable as well as the general challenges accessibility to data and the need for companies and institutions to share data for the public good whilst protecting the data subjects right to privacy.

[6] Researches the use of a popular Chinese micro-blog called ‘Sina’ to predict the health status of citizens in a region of Beijing. Users of the micro-blog can make posts of 140 characters or less. Data was accessible via an API and the researchers were able to limit the data source to nearby users. They categorised 5000 posts into sick and not-sick for their training and test set. Thus, adopting a supervised learning approach. Programming was carried out in Python. Pre-processing included word-weighting utilising Boolean weighting, term frequency weighting, inverted document frequency weighting and term frequency-inverted document weighting frequency. A classification approach utilising K-Nearest Neighbour and Support Vector Machine algorithms were tested. Their algorithm was able to predict an outbreak of influenza 5-days before the Chinese national influenza centre.

[7] Demonstrates the used of multiple Big Data sources to perform a retrospective study of prediction of outbreaks of COVID-19. Taking a two-step approach, first step being a lag correlation analysis to find the number of days ahead to make an accurate prediction. Secondly to apply Machine Learning algorithms. The methodology applied was as follows: four Big Data sources, social media, web-search, air quality and daily COVID-19 new cases were pre-processed, underwent Log Correlation Analysis, and then fed into Machine Learning and Deep Learning based prediction models. The pre-processed web-based social media and search data being sorted into data collected during lockdown/ unlock phases and that collected outside of nationwide lockdowns. A regression approach was taken for the Machine Learning approach, utilising the Support Vector Regression and polynomial algorithms. Several Deep Learning algorithms were used: multi-layer perceptron, ElasticNet and deep neural network. They conclude that daily new cases could be accurately predicted 70 – 100 days ahead with the inclusion of air quality data improving the performance of the models.

[8] The method proposed by these researchers is again a hybrid-model including data source from the United States’ Centre of Disease Control [CDC] and flu-related data sourced from Twitter. The objective of the paper is to predict future influenza activities, to do so more accurately than current methods, and to combine the real-time feed from Twitter with historical data from the CDC. Using the most recent five weeks of data from both sources to produce a model that accurately predict influenza 2-3 weeks ahead. Their model improved upon the predictions of the 2013 version of Google Flu Trends.

[9] This study, focusing on sub-Saharan Africa, identifies two important trends. Firstly, that searches on the world-wide-web for health-related information can be helpful in monitoring disease trends, and the terms of the searches are specific to sub-Saharan Africa, insofar as they may not be predictive of influenza trends in other countries. Secondly, they highlighted the need for greater public health education regarding diseases to aid the citizens in conducting safe and productive internet searches. The researchers conducted a hybrid approach using data from Cameroon’s surveillance system ILI as well as Google Search data. They evaluate several multivariate regression models, the Random Forrest ensemble model and ARIMA.

[10] Propose Hadoop as the pre-processing platform for Social Media Big Data Analytics. Data is sourced from Twitter, and whereas Relational Database Management Systems are traditionally used the authors propose using Hadoop’s Map Reduce Model to transform the data into key-value pairs to increase the data volume, processing capability, deal with structured and unstructured data, for producing ad hoc reports, fault tolerance, all in a more cost-effective platform. Apache Hive is used for querying the data and real-time tweets are streamed from Twitter using an API. Their methodology aimed to analyse Tweets within the minimum time.

# **Healthcare Big Data analytics**

[11] Investigate Big Data Analytics in healthcare, studying its unique characteristics, different analytics phases, platforms, and tools used, challenges and their conclusions. Ending with a discussion of its further evolution. The unique characteristics, commonly referred to as V’s, describe the data in four dimensions: Volume, Velocity, Variety and Veracity. Volume relates to the amount of data. Velocity relates to the rate at which the data is created. Variety relates to the state of the data: structured, unstructured, and semi-structured. Finally, Veracity relates to the trustworthiness that the data is suitable and valuable for the purpose in which it will be used. They conduct a literature review of the existing systems in which data is stored and used. Their analysis summaries to exiting system and proposes improvements. The weakness of the existing systems being described as its inability to support live streaming, limiting the availability of the data. Their proposal lays out a technology for improving data availability for its present architecture using the MapReduce as well as proposing the use of The Spark and Kafka technology for a new big data framework.

[12] Discuss the processes that make Big Data in health care different from nominal data. They break it down in terms of features [Structured Features, Linked Data, Multisource and Multiview Data, Streaming Data and Features, Scalability and Stability], challenges in Big Data representation and reduction, then discuss the merits of a variety of potential classifications of Machine Learning algorithms. They discuss the variety of data stored in electronic health records includes health data, demographics, medical history and more, that can be collected using techniques such as interviews, sensors, laboratory analysis and IOT devices. Identifying that before Machine Learning techniques are applied practical issues need to be overcome in how this data is networked, secured, and distributed. The data pre-processing steps include data organisation and structuring, data cleaning and reduction, data integration and processing. Thereafter, selecting, tuning, and evaluating a variety of algorithms can occur.

# **Demographic Big Data**

[13] Investigate the use patient level data using the New York State Department of Health Statewide Planning and Research Cooperation System. Their method combines this data with high resolution spatial and demographic analysis techniques to assess census-tract level variation in breast cancer screening and diagnosis rates. Geo-spatial analysis was used using the addresses of target patients and population data from the American Community Survey of the US census survey. They found significant differences in the rate of Breast Cancer screening by race and discussed variances is diagnosis and mortality rates across different race and ethnic groups. Concluding that incorporating demographics into predictive models can assist in a targeted approach to help address these disparities. This is of interest as a comprehensive model of predicting COVID-19 cases could potentially account for demographic factors that influence the transmission of the virus,

[14] This study seeks to implement baseline calculations for the prediction of demographic markers to identify demographic markers in photos, biographies and tweets. Taking a bag-of-words approach to feature selection and then evaluating several machine learning models. In their model they find it easier to predict age and gender than age and location. Even so, they claim their approach has advantages, namely addressing bias present in hard-coded dataset and eliminating the effort required to hard-code datasets for a Corpus. Such a method could be helpful for aligning information contained within Social Media posts with demographic data gathered from Census data, or example to compare the extent to which the social media posts represent the overall population in terms of age, gender, ethnicity, race, etc.

# **Conclusion**

<< This is a critical part of the literature review. Briefly state again the overall context of the work. Briefly state the literature review process including technical core areas and subcategories. Compare and contrast the relevance of the technologies under discussion. Compare and contrast the potential for existing technologies to address the overall research challenge. Identify limitations in existing approaches. Discuss future work which can enhance the existing state of the art in order to address the overall research challenge>>

# **Works Cited**

|  |  |
| --- | --- |
| [1] | Oracle, “What is Big Data,” [Online]. Available: https://www.oracle.com/big-data/what-is-big-data/. [Accessed 31 01 2022]. |
| [2] | L. Samaras, E. Garcia-Barriocanal and M.-A. Sicilia, “Comparing Socila media and Google to detect and predict severe epidemics,” Springer Nature, 2020. |
| [3] | A. Parwez, M. Abulaish and Jahiruddin, “A Social Media Time-Series Data Analytics Approach for Digital Epidemiology,” in *2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, 2020. |
| [4] | L. Albancdes, B. A. Bungar, J. P. Patio, R. J. M. Sevilla and D. Acula, “Application of C5.0 Algorithm to Flu Preduction Using Twitter Data,” in *2018 International Conference on Platform Technology and Service (Platcon)*, Manila, 2018. |
| [5] | D. Lazer and R. Kennedy, “What Can We Learn from the Epic Failure of Google Flu Trends,” Wired, 1 October 2015. [Online]. Available: https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/. [Accessed 02 2 2022]. |
| [6] | N. Yang, X. Cui, C. Hu, W. Zhu and C. Yang, “Chinese Social Media Analysis for Disease Surveillance,” in *2014 International Conference on Identification, Information and Knowledge in the Internet of Things*, Wuhan, 2014. |
| [7] | 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, 2020. |
| [8] | K. Lee, A. Agrawal and A. Choudhary, “Forecasting Influenza Levels using Real-Time Social Media Streams,” in *IEEE International Conference on Healthcare Informatics*, Evanston, 2017. |
| [9] | E. Nsoesie, O. Oladeji, A. Abah Abah and M. L. Ndeffo-Mhab, “Forecasting influenza-like illness trends in Cameroon using Google Search Data,” Nature, 2021. |
| [10] | F. Shaikh, F. Rangrez, A. Khan and U. Shaikh, “Social Media Analytics Based on Big Data,” in *International Conference on Intelligent Computing and Control*, Mumbai, 2017. |
| [11] | G. Sasubilli and A. Kumar, “Machine learning and Big Data Implementation on Health Care data,” in *International Conference on Intelligent Computing and Control Systems*, 2020. |
| [12] | P. Saranya and P. Asha, “Survey on Big Data Analytics in Health Care,” in *Smart Systems and Inventive Technology*, 2019. |
| [13] | F. Murphy, K. Abel-Hart and F. Wang, “A Fine-grain Geospatial and Demographic Analysis of Breat Cancern Patterns in New York State,” in *IEEE International Conference on Big Data*, 2021. |
| [14] | J. Radford, L. Horgan and D. Lazer, “Baselines for Demographic Inference on a New Gold Standard Twitter Corpus,” in *IEEE International Conference on Big Data*, Boston, 2017. |