Assessment 1: Research Proposal

Discovering how Publicly Funded Data is used for Public Good

BIRMINGHAM CITY University

Name:

University ID:

Word Count:

Date:

Supervisor:

Table of Contents

Table of Figures

Abstract

Access and sharing of data are essential for the conduct and development of technological know-how. This thesis argues that publicly funded research information needs to be brazenly available to the most extent viable for public suitable. To seize upon advancements of cyberinfrastructure and the explosion of facts in quite a number of scientific disciplines, this access to and sharing of publicly funded records should be superior inside an international framework, past technological solutions. Data collection is a major bottleneck in machine learning and a lively study subject matter in a couple of groups. Interestingly, current studies in information series come no longer best from the machine learning to know, natural language, and computer vision groups, however additionally from the records management network due to the importance of coping with large amounts of records. Machine learning gives computer systems the potential to research from data without being explicitly programmed. Due to its wonderful prediction abilities, it has currently received traction in economics, statistics, and social sciences. In this thesis, I will explore how gadget getting to know can assist to clear up such and other prediction problems in public coverage making and what demanding situations it faces. This proposed work will be an efficient, reliable, and robust system that is based on a machine-learning algorithm. Keeping in view the projection of synthetic intelligence the researcher has taken up the system gaining knowledge of as a method to inventing based mostly on Natural language processing for purchasing more appropriate quit results. It will explore the use of machine learning primarily based on Nature Processing Language models and their performance to evaluate the accuracy with a one-of-a-kind method for Publicly Funded Data that will use for Public Good.

Keywords: Machine learning, predictive modeling, prediction, public policy, social good.

Introduction

New Information and communication technology (ICTs), now insignificant use at some stage in all studies disciplines, have significantly aided this device of loose trade and feature unfolded new avenues for collaboration and sharing. The development of science, but, depends on greater than simply technology. Research policies, practices, guide structures, and cultural values all affect the character of discoveries, the charge at which they're made, and the diploma to which they may be made on hand and used.

In recent years, the UK Government has improved public spending on basic research in universities and studies laboratories. Yet scientists and research funding companies continuously argue that we need to spend even extra public money on studies. Government, but, faces several competing demands for public investment. For a number of those, inclusive of fitness, and training, the economic and different benefits perhaps seem to be more on the spot and obvious. Nevertheless, as we shall demonstrate, there's now an extensive frame of research on the economic and social benefits of publicly funded simple studies, and we will begin to see some of the benefits that accrue from publicly funded research (Bloom et al, 2019).

The fast development in IT over the most recent twenty years has prompted development in the measure of data accessible through computing devices. The history of the Internet is not too old for the world, it's been 30 to 40 decades since the revolutionaries of the Internet captures the world's attention and makes it a global village. Long-distance interaction with friends and families has been a great concern for people for centuries. The power of computers and the Internet has created new fields of utility for no longer only the outcomes of studies, however the resources of studies: the bottom cloth of research facts. Moreover, research records, in digital shape, are an increasing number of being utilized in studies endeavors past the original undertaking for which they had been amassed, in other research fields, and the enterprise (Marginson, 2018).

Effective access to research information, responsibly and efficiently, is needed to take full gain of the new possibilities and advantages supplied by ICTs. Nowadays, information technology occupying modernization changes all business methods providing simulation techniques to figure out in best achievement. IoT creates big data that send manipulate, manage, data that can make the decision automatically. IoT now considers the information technology development technique that combines the real-world with the imaginary world of information technology, new demands for the companies to develop complex systems instead of the classical system that integrates our physical world combine with the imaginary world of information technology. Data produced by the internet of things process can enhance the output of the company, give an automatic solution, and decrease nonessential costs (Ammirato et al, 2019).

Scientific databases are rapidly turning into an essential part of the infrastructure of the global technology machine. The worldwide Human Genome Project is however one good instance of a massive-scale research enterprise wherein an openly on hand facts repository is getting used efficiently with the aid of many extraordinary researchers, all over the international, for specific functions and in distinctive contexts. Many different examples, related to research undertakings both massive and small, are without problems available (Chimienti, 2019).

Although publicly funded research yields advantages well past production of the latest know-how, this remains one of the foremost exploitation channels for changing the result of research into improvements that gain society. Publicly funded studies and researchers additionally make contributions to the economy with the aid of assisting industry and others to solve issues (Ledyard, 2020). Many corporations in technologically stressful industries face complicated technological challenges, the solution of which regularly includes combining a spread of technologies in complex methods. Publicly supported research offers an intensive pool of resources for solving problems from which those corporations may additionally draw. In precise, graduates trained by way of researchers in technology and engineering are regularly very adept at tackling and fixing strange troubles. Studies consisting of the ones with the aid of (Park et al, 2017) observed that corporations' advantage notably from the recruitment of trained trouble-solvers together with these.

Many of the advantages that waft from publicly-funded natural technology and engineering are reflected in the social sciences (Smeaton, 2017). The social sciences have supplied the idea for such public goods as countrywide statistics, censuses, monetary models, and huge elements of the toolbox of the cutting-edge management of economies, all of which contribute in essential approaches to the innovation manner. Indeed, the whole manner in which society perspectives itself and tries to expand regulations for the development of society is inextricably linked to traits inside the social sciences (Archibugi and Filippetti, 2018).

There are large reasons information collection has currently become important trouble. First, as gadget getting to know is becoming greater broadly used, we are seeing new applications that don't necessarily have enough categorized information. Second, not like traditional machine studying strategies robotically generate functions, which saves feature engineering fees, however in go back can also require larger quantities of labeled statistics (Marginson, 2018). Aim is to convey the two fields closer collectively as maximum public coverage makers likely do not even realize that they face prediction troubles that machine learning can help fix. After a creation to prediction troubles, I supply an overview of the way device studying works and explain for below what instances machine learning can be used for statistics-driven predictive modeling for the social good.

Machine Learning is a subfield of artificial intelligence where software programs become able to classify and predict results accurately without programming them explicitly (Das et al, 2015). The learning process of machine learning software modules involves providing some data for those models, allowing those models to look for patterns into data, and make better decisions in the future based on the data provided. The main aim of machine learning is to allow software programs to be learned directly from provided data and adjust their results according to this data without the aid or interference of humans. Machine learning algorithms are often categorized into four categories: Supervised Algorithms (SA), Unsupervised Algorithms (UA), Semi-Supervised Algorithms (SSA), and Reinforcement Algorithms (RA) (Khan et al, 2020). In supervised algorithms data that is used to training modules (training set) is provided with labels. However, in unsupervised algorithms training data has no labels. In semi-supervised algorithms, both labeled and unlabeled data are used for training modules. In reinforcement algorithms modules are trained from the environment through a trial and error process.

Deep extreme machine learning is a subfield of machine learning. Deep extreme machine-learning generally uses sequences of several layers to accomplish the feature extraction and classification tasks. Layers used in deep extreme machine learning are connected in a cascade manner so that the output of each layer is connected to the input of the following layer (Nguyen et al, 2019). With deep extreme machine learning, software modules can be learned and trained to accomplish classification and prediction tasks from images, sounds, videos, or text data. The performance and accuracy of deep extreme machine learning models can be very excellent and exceed human beings' performance. Deep extreme machine learning models are trained to accomplish classification or regression tasks by using a large number of datasets (data with labels) and powerful neural network structures.

An artificial neural network is another subfield of machine learning that is briefly stimulated by the human neural network and it employs different neurons to perform amassed tasks. This technique has high-level accuracy because it leans towards the updates itself without human interaction. The forward technique of neural network process different instructions, and if it finds errors, it back propagates automatically, and improves back neurons to process data with high accuracy. The study reveals that it was up to mark the level of accuracy in the diagnosing system of medical fields (Soltani and Jafarian, 2016).

The advent of technology has made it easier for humans to automate the system and report analysis. However, Machine Learning is only as capable as the data and the resources it has. It cannot expand and learn at will. But the capacity of digesting huge amounts of data and statistically analyzing to provide accurate predictions in the blink of an eye is what draws the researchers to using ML for this purpose. This research aims to render out the limits of Machine Learning in achieving public funding data for the public good using Machine Learning. The output of this research will be the experimental proof and analysis of how much a Machine Learning algorithm can perform and evaluate accuracy for Publicly Funded Data. This proposed work will be an efficient, reliable, and robust system that is based on a machine-learning algorithm. Keeping in view the projection of synthetic intelligence the researcher has taken up the system gaining knowledge of as a method to inventing based mostly on Natural language processing for purchasing more appropriate quit results. It will explore the use of machine learning primarily based on Nature Processing Language models and their performance to evaluate the accuracy with a one-of-a-kind method for Publicly Funded Data that will use for Public Good.

Aim(s)

To build a Machine Learning Model that identifies the mention of publicly funded datasets within scientific publications.

Objectives

- Proposal
- Data Pre-processing
- Natural Language Processing
- Machine Learning Models
- Performance Metrics

Research Question

- Is the data collected publicly used to serve science and society?

Ethical Considerations

Ethics is a complicated subject that has only become more prominent during the advent of Big Data. The UK Data Service department also provides guidelines for ethical research with specific relation to Big Data. These guidelines will form the basis for this report's ethical approach. Some of the concerns that will be addressed are:

- Maintaining confidentiality in line with Birmingham City University (BCU) and DC guidelines.
- Anonymizing information that violates group privacy.
- Ensuring transparency in reasons for data collection.
- Ensuring data is only used for the direct purpose it has been requested
- Referencing sources for all information used within the research project.
- Ensuring all data is stored in the correct location. DC information must remain on DC servers.

As this project encounters any further ethical concerns these will be met within the recommended UK guidelines and with the advice of BCU and DC supervising members.

Literature Review

The process of reviewing, studying, and understanding research was done by peer researchers in the same domain through their research papers is known as Literature Review (LR). A strong LR provides validity to the integrity of the research using proven, published facts. The papers selected for LR in this research are summarized as follows.

The literature identifies varieties of studies-primarily based understanding – codified and tacit. Codified expertise comes in a written form and is the greater visibility of the two. Tacit information refers back to the skills, expertise, and revel in added to any assignment by using those sporting it out and is for this reason embodied in human beings, who bring it around with them once they pass.

(Sulkunen et al, 2018) explained the conventional justification for public funding of fundamental studies is based broadly speaking in this 'Channel 1' exploitation mechanism. According to the 'marketplace failure' cause for public funding of studies, simple studies expand the pool of medical expertise available to corporations and other 'customers' who can draw in this freely in their technological sports. However, this argument underplays or ignores as a minimum of three matters.

(Poole and Garwood, 2018) there is evidence that corporations draw extensively on new scientific ideas. Publicly funded fundamental research frequently stimulates and complements R&D done by way of corporations, as well as expanding the range of technologically exploitable opportunities. It explores this issue of how publicly funded studies can stimulate extra R&D efforts by using enterprise, for example, encouraging firms to interact in greater collaboration in R&D projects. One observes reviewed within the OECD file indicated that existing partnerships had been intensified and new ones initiated because of government investment. Another confirmed that many consortia and joint tasks had been shaped at once due to the authority's investment, and that collaboration regularly persevered beyond the participation in a central authority-funded project (Hazelkorn and Gibson, 2019).

(Venturini et al, 2018) there had been numerous attempts to degree the monetary impact of publicly funded research and development (R&D), all of which show a big high-quality contribution to monetary growth. Another recent take a look at unearths that R&D funding, as a whole, and better schooling R&D investment, especially, are undoubtedly related to innovation and economic growth in peripheral regions of the EU, even though the power of this association varies with the socio-economic characteristics of each area.

(Park et al, 2017) explained the Econometric research of the benefits from studies normally contains the statistical analysis of massive databases. Early paintings focused on demonstrating that an enormous share of monetary increase should be attributed to technological exchange in preference to changes in labor or capital. The boom in peripheral areas of the EU, even though the power of this association varies with the socio-monetary characteristics of each.

(Perwej et al, 2019) carved that more than 34 billion devices will be by the end of 2021. IoT will give rare classifying for the reorganization of devices. A Big data transfer through the internet of things devices, a lot of technologies works for conversion of large data. IoT has a functionality of plate form, the IoT PLATFORM is an important element for large communication of data as well as large components for communication of data. IoT platforms act as a middleware for connecting two devices, IoT platforms provide built-in features that are easy to understand also provide a faster way to develop applications. This platform acts as a middleware between applications and remotely connected devices, control all the intercommunication between physical devices and software. In this paper, the internet of things creates new devices to make more efficient in utilization of energy with a low cost like smarty meter, recovery and preserving, gather action data easy as well as electricity pole watching these technologies creates innovation in many fields of energy. (King and Ballantyne, 2019) implemented the IoT in health care departments: laboratories test outcomes, human-based data, medical treatment history as well as other facilities provided by these technologies. On the internet of thing in a smart home domain: needs all appliances can be operated with single gadget this can be done with the internet of things. Many stakeholders deepened on the real smart city system there are some manufactures: device creators, end-user, network service providers, administration purpose as well as many kinds of services provider work for the smart city domain. Now we can say that IoT applications use in every field of life.

(Howe et al, 2017) proposed the scientific studies is frequently visible as a way of spurring the boom of latest corporations. Researchers and students can spin out of universities to exploit new ideas and technology by using organizing start-up groups, accordingly transferring talents, tacit know-how, and problem-solving skills, and so on from academia to an industrial surrounding. While the focal point of this review thus far has been on the medical and technical inputs to innovation, few problems can be solved on the premise of scientific and technological information alone. In much new technologies, innovators additionally face non-technical demanding situations that involve social choices. For instance, groups inside the fitness area ought to cope with large regulatory hurdles.

Project Timeline

Research projects are random and time-bound and the ability to meet a deadline is key to success. A project timeline lays out key project deliverables and the scope of their completion. This research project identifies time as its key resource. By efficiently allocating time to various tasks resource overload is minimized. Preventing resource overload minimizes the risk of quality decreasing. The Gantt chart is identified as a strong tool for time management. The Gantt chart designed for this project is laid out below in Figure 1.

Figure 1**:** Project Timeline

Methodology

The dataset initializes a problem that requires the prowess of Natural Language Processing tools to get through. The processes that go into the Processing Model will be Literal Matching and Ken Matching by creating a Knowledge Bank.

The research investigating the differing aspects, may be proposed that the research methodology would be a diverse one, specified, and altered towards the various factors of research objectives. Within the consideration of the theoretical aspects of the research, it could be established that the research would significantly be focused upon secondary research as to identifying the previous proposed work and factors in relevance to the research and its objectives. However, moving on to the constructive and practical aspects of the research, it would be essential to shift the methodology in a manner where statistics and quantitative procedures would be required. Because the result of quantitative data is more effective and arranged. Data will collect from a different sources and some sites of data science and the basic purpose of collecting data is to the enhancement of publicly-funded data for the public good in the era of networking and Information Technology. Collected data will be used through some techniques or algorithms on behalf of machine-based learning using Natural Processing Language for getting a more efficient and appropriate results.

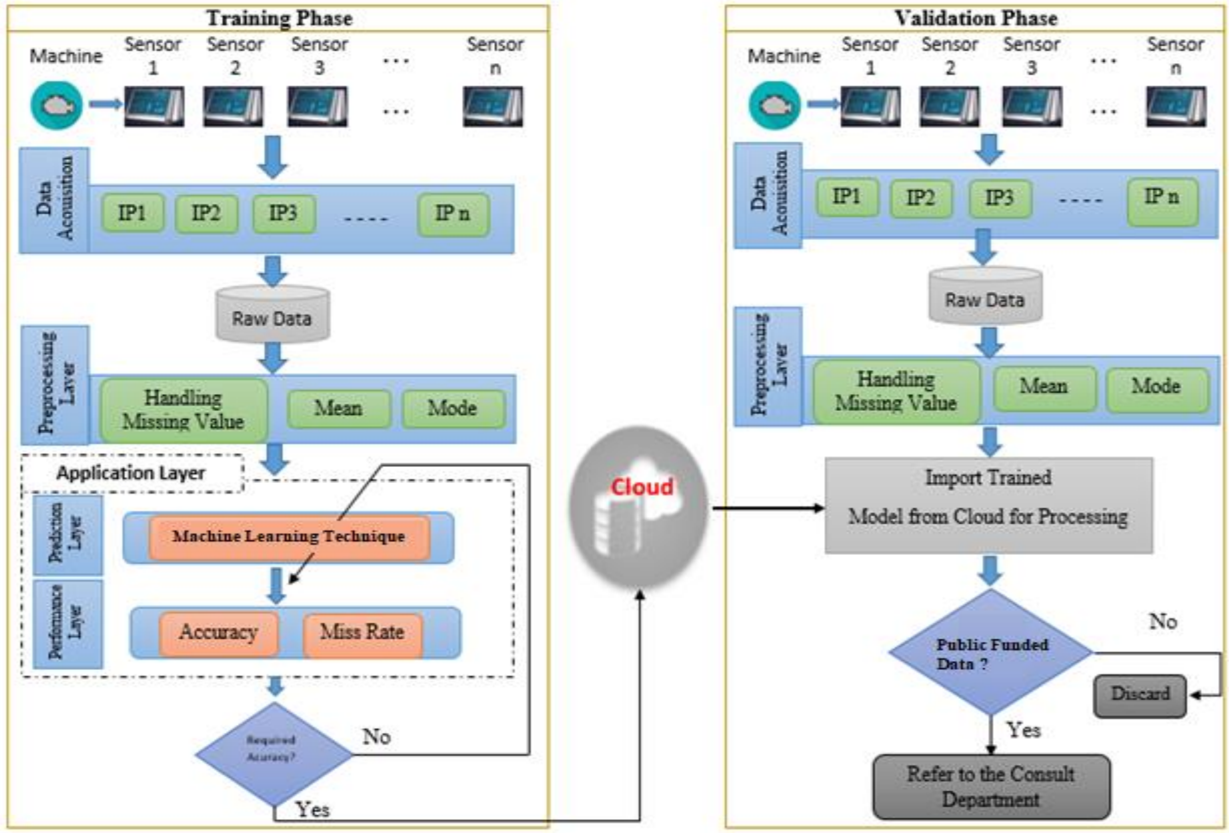

Figure 2: Proposed Methodology for Public Funded Data

Project Evaluation

This research gives immense improvement to build a Machine Learning Model that identifies the mention of publicly funded datasets within scientific publications with the involvement of machine learning with the integration of Natural language processing that is undermined and is treated as underdogs while gaining significant insight into the differences between the state-of-the-art and the novel approach.

Conclusion

The current debate on open access to the results of publicly funded research focuses primarily on research in the field of natural sciences, technology, and medicine, but arguments for open access can be applied also to research in the legal field. Quantify the economic and social benefits from publicly funded research is beset by way of troubles. This report suggests how the benefits from publicly funded research come in a ramification of paperwork, flowing through a variety of channels and over differing timescales. Government coverage wishes to mirror this fundamental point and to find powerful approaches of influencing the taking into consideration agencies accordingly. It will explore the use of machine learning primarily based on Nature Processing Language models and their performance to evaluate the accuracy with a one-of-a-kind method for Publicly Funded Data that will use for Public Good.

References

- Ammirato, S., Sofo, F., Felicetti, A.M. and Raso, C., 2019. A methodology to support the adoption of IoT innovation and its application to the Italian bank branch security context. *European Journal of Innovation Management*.
- Archibugi, D. and Filippetti, A., 2018. The retreat of public research and its adverse consequences on innovation. *Technological Forecasting and Social Change*, *127*, pp.97-111.
- Bloom, N., Van Reenen, J. and Williams, H., 2019. A toolkit of policies to promote innovation. *Journal of Economic Perspectives*, *33*(3), pp.163-84.
- Chimienti, A., 2019. Saving The National Endowment for the Arts: Why Publicly Funded Art Serves as a Public Good.
- Das, S., Dey, A., Pal, A. and Roy, N., 2015. Applications of artificial intelligence in machine learning: review and prospect. *International Journal of Computer Applications*, *115*(9).
- Hazelkorn, E. and Gibson, A., 2019. Public goods and public policy: what is public good, and who and what decides?. *Higher Education*, *78*(2), pp.257-271.
- Howe, A., Mathie, E., Munday, D., Cowe, M., Goodman, C., Keenan, J., Kendall, S., Poland, F., Staniszewska, S. and Wilson, P., 2017. Learning to work together–lessons from a reflective analysis of a research project on public involvement. *Research involvement and engagement*, *3*(1), pp.1-12.
- Khan, F., Khan, M.A., Abbas, S., Athar, A., Siddiqui, S.Y., Khan, A.H., Saeed, M.A. and Hussain, M., 2020. Cloud-based breast cancer prediction empowered with soft computing approaches. *Journal of Healthcare Engineering*, *2020*.
- King, M. and Ballantyne, A., 2019. Donor-funded research: permissible, not perfect. *Journal of medical ethics*, *45*(1), pp.36-40.
- Ledyard, J.O., 2020. *2. Public goods: A survey of experimental research* (pp. 111-194). Princeton University Press.
- Marginson, S., 2018. Public/private in higher education: A synthesis of economic and political approaches. *Studies in Higher Education*, *43*(2), pp.322-337.
- Nguyen, G., Dlugolinsky, S., Bobák, M., Tran, V., García, Á. L., Heredia, I., ... & Hluchý, L. (2019). Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey. *Artificial Intelligence Review*, *52*(1), 77-124.
- Park, S., Stone, S.I. and Holloway, S.D., 2017. School-based parental involvement as a predictor of achievement and school learning environment: An elementary school-level analysis. *Children and Youth Services Review*, *82*, pp.195-206.
- Perwej, Y., Haq, K., Parwej, F., Mumdouh, M. and Hassan, M., 2019. The internet of things (IoT) and its application domains. *International Journal of Computer Applications*, *975*(8887), p.182.
- Poole, A.H. and Garwood, D.A., 2018. Interdisciplinary scholarly collaboration in data-intensive, public-funded, international digital humanities project work. *Library & Information Science Research*, *40*(3-4), pp.184-193.
- Smeaton, A.F., 2017. Open Access to Publications and to Data from Publicly Funded Research. Ireland and the World. *Education Matters Yearbook*, *2017*, pp.402-406.
- Soltani, Z., & Jafarian, A. (2016). A new artificial neural networks approach for diagnosing diabetes disease type II. *Int J Adv Comput Sci Appl*, *7*, 89-94.
- Sulkunen, P., Babor, T.F., Ornberg, J.C., Egerer, M., Hellman, M., Livingstone, C., Marionneau, V., Nikkinen, J., Orford, J., Room, R. and Rossow, I., 2018. Setting limits: Gambling, science and public policy.
- Venturini, T., Munk, A. and Meunier, A., 2018. Data-Sprinting: a Public Approach to Digital Research.