|  |  |
| --- | --- |
| MASTERS INDIVIDUAL PROJECT PROPOSAL(60 credits)  Supervision Acceptance Form | |
| Student Name | *ABC* |
| Project Title | *Classifying fake news articles using natural language processing to identify in article as a supervised lear* |
| Preferred Supervisor |  |
| Project Aims | Problem Statement Difference between fake and real news is hard to find for a normal human being. Therefore, the problem that needs to be solved is to identify the fake news using data science application and its methodology. Project Aims The focus of the problem is to design the data science tools using various data related to real and fake news. Machine learning capability will automatically upgrade itself when there is fake news detected. Designing a flawless machine learning through data science has been done is the project.***LSTM*** networks are very good at holding long term memories or in other words, the prediction of nth sample in sequence of test samples can be influenced by an input that was given many times steps before. The long short type memory may or may not be retained by the network depending upon the data. Sherstinsky (2020) has said that long term dependencies of the network are processed by its Gating mechanisms. The network can store or release memory on the go through the gating mechanism. Thus LSTM is a good choice for such sequences which have long term dependencies in it. Therefore LSTM is used over other existing models. Project Objectives  * To understand if a data science application uses proper sets of data to analyse and gain information for analysis * To analyse the data and make decisions depending on the data * To design various tools related to data science according to the requirements of the project  Research Question  * How a data science application uses proper sets of data to analyse and gain information for analysis? * How to analyse provided the data and make decisions depending on the data? * What is the way to design various tools related to data science according to the requirements of the project?  Success Measurement Data science application is smart and very reliable to do any task. The designed application will easily identify the difference between real and fake news. Machine learning will be used to identify the difference between real and fake news. |
| Rationale | The issue is to counter the issue, machine learning needs to be implemented to identify fake news.  This is an issue becausefake news is circulated through online media platform and social media.  This is an issue now becausethe fake news is harming people’s lives in various ways, which are causing serious damage to their careers, and life as well.  The research will shed light onthe methodologyto identify and implement a proper way to encounter the fake news issue. |
| Planning | Justification of planning tool Data science methodology uses various types of data from the user and analyses it with various references. Partitioning the data into various datasets to and compare, it with the various hypotheses is the key planning of the method. After partitioning the data sets, fresh data need to be inserted to compare with previous data to enhance the learning capability. Data science application analyses and evaluates the data with various algorithms and methods (Northeastern, 2020).  Machine learning uses ***python*** language to implement data science analytics in its system. In the implementation of data analysis SQL database, data robot, apache-spark software needs to be used.  ***Risk analysis***  Various risks is also involved with the data analysis. Proper use of data evaluation respect with references needs to be taken under consideration. While data analysis, there are some evaluation factors that python does not recognize which causes the issue related to the data clarification. Sometimes it becomes difficult to identify the original source of the data, which leads to the issue of data originality (Northeastern, 2020). For this reason, the implementation of data science analytics needs to be more enhanced and sophisticated regarding data analysis using various database and language.  ***Timeline***        **Figure 2: Timeline** |
| Method | Research Method Aim of the project is to identify the fake news by analysis of the quality and structure of data. The main method used to analyse the data is to implement and design the codes using python language. Identification and evaluation of data need to b practised before implementing it into the real world. First, some amount of fake news and real news needs to be inserted in the database to help in the learning process (Alonso-Fernández*et al.,* 2019). After learning, the structure of the data system will easily identify the difference between real and fake news. Justification of tools Designing the data analytics application main requirement is the language through which the system will perform its task. In this chosen project, machine learning has been used to analyze the data and taking the decision related to it. As discussed by Sharma (2018), machine learning is a very complex process and it requires a complex methodology to implement a decision-making tool. Python is the most readable and easily understandable code while comparing with the other codes in the area of machine learning.  ***Apache spark*** has been used to design the data analytic engine. It the most used data analytic engine all over the world. This engine accompanies various APIs to analyse the data and comes up with a proper evaluation.  The most important part of the project is the database, which will be used to store the required data. ***MySQL*** is the vastly used database services for data analytics. MySQL offers a synchronized data structure related to any project (Roy et al., 2017). The data stored online and can be accessed from anywhere. MySQL database can be equipped with python and make a data analytics application easily. Application of appropriate methods First, the system needs to be designed according to the data structure related to the fake and real news. In the designing process, the implemented codes need to be synchronized with the data structure. In the beginning, the system will learn the difference between fake and real news through the inserted data (Polonsky*et al.,* 2019). After learning the difference, the system will learn to make decisions according to the provided data.Collecting, analysing and visualizing of fake news is observed by fake news tracker tools. The fake database showing that there are no news channel’s names are showing where as in the original dataset every channel has individual headquarters. Manipulating the idea of dataset fake channels are using a news portal which is not yet registered. Therefore with the help of the original dataset one can compare them and identify them specifically. The complexity of the project The project complexity is high depending on the project objectives and project development cycle (Cao, 2017). The project involves machine learning approach in the form of deep learning. The project will utilize an arbitrary datasets for detecting the fake news. The dataset cannot be distributed completely because of twitter privacy policy. Developing machine learning programs can identify an article whether it fake or not. The datasets are collected from different sources. This dataset contains different types of articles on different topics. Majority of the articles focus on world news and politics therefore the fake news articles are collected from unreliable websites (Salem*et al*. 2019). The data collected were fresh and processed however the punctuations and mistakes that existed in the fake news were kept in the text. |
| Conclusions | Identify fake news through the data analytic has been the new addition in data science. In the concept of emerging technology, it becomes very difficult to identify the authenticity of the data circulated online. In this procedure, data science has been the helping tool for various resources. |
| Project Evaluation | Evaluation of the process The main highlighting part of the project is to identify the fake news using data science application. As argued by Sharma (2018), in the use of identifying fake news, the system needs to be designed to analyse the source of the data and check its authentication. This process will include machine-learning techniques to evaluate the data and make the decision. Links to objectives The objective of the project is to design a system, which will be efficient to identify the authenticity of any news circulating in the online platform.To design the system various source of data has been analysed and compared with various references. Link to literature review Data science technology has been developing through many generations and improving day by day. For designing the system, data science has played a vital role. Data from various sources have been analysed through the machine learning process. This method has helped to learn the system to identify the data structure. There is some bottleneck related to the implementation of machine learning has been highlighted. |
| References | References **Books**  Sharma, M. (2018). *Data Science Analytics and Applications*. Singapore: Springer  **Journals**  Alonso-Fernández, C., Calvo-Morata, A., Freire, M., Martínez-Ortiz, I., &Fernández-Manjón, B. (2019). Applications of data science to game learning analytics data: A systematic literature review. *Computers & Education*, *141*, 103612.  Cao, L. (2017). Data science: a comprehensive overview. *ACM Computing Surveys (CSUR)*, *50*(3), 1-42.  Roy, S., Ray, R., Roy, A., Sinha, S., Mukherjee, G., Pyne, S., ...&Hazra, S. (2017, August). IoT, big data science & analytics, cloud computing and mobile app based hybrid system for smart agriculture. In *2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON)* (pp. 303-304). IEEE.  Salem, F. K. A., Al Feel, R., Elbassuoni, S., Jaber, M., & Farah, M. (2019, July). Fa-kes: A fake news dataset around the syrian war. *In Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 13, pp. 573-582).  Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena*, *404*, 132306.  Zhou, X., Zafarani, R., Shu, K., & Liu, H. (2019, January). Fake news: Fundamental theories, detection strategies and challenges. In *Proceedings of the twelfth ACM international conference on web search and data mining* (pp. 836-837).  **Online Articles**  Polonsky, J. A., Baidjoe, A., Kamvar, Z. N., Cori, A., Durski, K., Edmunds, W. J., ...& de Waroux, O. L. P. (2019). Outbreak analytics: a developing data science for informing the response to emerging pathogens. *Philosophical Transactions of the Royal Society B*, *374*(1776), 20180276. Retrieved from < https://doi.org/10.1098/rstb.2018.0276> Retrieved on [05/12/2020]  **Websites**  Northeastern.edu, 2020 *DATA ANALYTICS VS. DATA SCIENCE: A BREAKDOWN* Retrieved from <<https://www.northeastern.edu/graduate/blog/data-analytics-vs-data-science/>>Retrieved on [05/12/2020] |
| Signoff | Project Supervisor SignatureDate |

Proceed to Project is subject to completing all taught units and recommendation by Board of Examiners. This form is NOT approval to proceed.

On completion please submit this form to the Programme Administrator as soon as possible. Supervisors will be allocated, by the Head of Academic Group and Programme Leader, as appropriate to the project subject, and on a first come first served basis.