# MSc Artificial Intelligence and Data Science

# Capstone Project Proposal

|  |  |  |
| --- | --- | --- |
| **Project Title** | | |
| **RiskTracker: A Text classification framework for news article tagging and risk type classification.** | | |
| **Apprentice Title** | **Apprentice First Name** | **Apprentice Last Name** |
| Mr | Dylan | Viswambaran |
| **Employer Name** | **Mentor/Line Manager Name** | **Mentor/Line Manager Email** |
| Bank of England | Paul Boyle | [Paul.Boyle@bankofengland.co.uk](mailto:Paul.Boyle@bankofengland.co.uk) |
| **Academic Supervisor Name** |  |  |
| TBC |  |  |

|  |  |
| --- | --- |
|  | You can make suitable adjustments to this template as required. The information in the blue boxes is here to help you understand what should be included in each section. |

## Project Background and Rationale

|  |  |
| --- | --- |
|  | Describe how this project came about, who is involved, the purpose of the project, and the expected benefits (value to the organisation). |

|  |
| --- |
| The rapid growth of digital media and the abundance of news articles makes it increasingly challenging to extract meaningful insights from the vast amount of news available. In light of this I’ve identified a need to develop an assisted process for supervisors that can tag and classify news articles for topics of interest such as specific risks (credit, market, liquidity, fraud, fines and so on). This project would involve collaboration between data scientists and domain specialists such as supervisors to enhance accuracy of the final product. This produce should enhance decision making by serving as an early warning indicator and allow decisions to be made sooner than currently possible |

## Project Scope

|  |  |
| --- | --- |
|  | Project scope defines the boundaries of a project. Think of the scope as an imaginary box that will enclose all the project elements/activities. It not only defines what you are doing (what goes into the project), but it sets limits for what will not be done as part of the project (what doesn’t fit in the project). Scope answers questions including what will be done, what won’t be done, and what the result will look like.  Refer to the competencies you will demonstrate in this project (in section 5). |

|  |
| --- |
| Data collection – gathering a diverse and comprehensive dataset of news article.  Model development – developing a robust text classification model which will be trained to tag and classify news articles.  Feature engineering – exploring various techniques which can include bag of words, word embeddings, transformers or zero shot classification.  Testing and validation – testing and validation will be conducted to ensure the accuracy and reliability of the text classifier. Involves evaluating performance against a separate test dataset and comparing the outputs with human labelled annotations.  The project scope will not include:   * News article collection infrastructure: the project will not address using any specific news article APIs or sources but will focus on the development of the text classification model which can then be applied to the preferred news article supplier when chosed. * Real time news monitoring: this pipeline will be built to focus on building a robust model to use rather than real time news monitoring. * Multi lingual support: the project will focus on news article written in English * Continuous model improvement: whilst the models will involve iterations of model training and evaluation it will not include continuous improvement upon deployment. These will take place in future phases of the project. |

## Aim(s) and Objectives

|  |  |
| --- | --- |
|  | Describe the overall aim of your project and the supporting objectives that support the development. For example, your aim might be:   * To develop a system that facilitates the information sharing between the marketing team and operations teams   With the following objectives:   * To identify the requirements of the marketing team and operations teams * To identify the appropriate system development methods and techniques that can be used for the project * To develop a system based on the agreed requirements and development methods * To implement and evaluate the developed system |

|  |
| --- |
| Aims:   * Develop a text classification model that can accurately tag and classify news articles based on risk types and associated terms of interest. * Enhance and improve decision making for supervisors by providing them with timely and relevant risk related information extracted from news articles * Act as an early warning indicator to proactively identify areas of interest particularly around specific risks reported in news articles.   Objectives:   * Collect and process a diverse and comprehensive dataset of news articles suitable to train and test the model * Explore and implement appropriate feature engineering techniques to extract relevant features from news text. * Train and evaluate the models using the collected dataset to optimise its performance and accuracy * Integrate the model into a user friendly application with supporting documentation on the model specifics. |

## Deliverable(s) and Success Criteria

|  |  |
| --- | --- |
|  | Describe your project deliverables and how they can be measured and considered achieved and successful. For example, the deliverables can be:   * Systems requirements documentation * A new system that meets the user requirements   The success criteria could be:   * Ability to allow both internal and external users to access the application without downloading any software * Ability to interface with the existing data warehouse application * Ability to incorporate automated routing and notifications based on existing business rules |

|  |
| --- |
| Text classification Model System: The primary deliverable of this project is a trained text classification model capable of accurately tagging and classifying news articles into risk types.  User interface: An interface to be created that receives news article inputs and outputs relevant risk tags. It should be user friendly, reliable and meet performance requirements.  Documentation: Comprehensive documentation including technical specifications, instructions and troubleshooting guides.  Success Criteria:  Accuracy of classification model: The success of the project will be measured by the accuracy of the resulting model tagging and classifying articles into the correct categories and on a consistent basis.  System performance and reliability: The performance and reliability of the model. It should be able to handle a reasonable volume of news articles within acceptable response times and demonstrate stability and resilience in real world usage.  Achievement of Project Objectives: This includes the development of the text classification model , the enhanced risk monitoring, streamlined information processing, improved/faster decision making and an establishment of a form of an early warning indicator. |

## Competencies to be Demonstrated

|  |  |
| --- | --- |
|  | The tables list the core and specialist competencies to be demonstrated via the project. In the ‘how will it be met?’ cells, please describe how the apprentice will meet each requirement through the proposed project. |

|  |  |
| --- | --- |
| Core competencies (applicable for all specialisms) | How will it be met? |
| K1 How to use AI and machine learning methodologies such as data-mining, supervised/unsupervised machine learning, natural language processing, machine vision to meet business objectives | The project will involve using AI and machine learning methodologies such as data mining, supervised/unsupervised machine learning, and natural language processing to develop the text classification model. These techniques will be applied to meet the business objective of accurately tagging and classifying news articles based on risk types. |
| K3 How to apply advanced statistical and mathematical methods to commercial projects | Advanced statistical and mathematical methods will be applied during the project to develop and evaluate the text classification model. These methods will help optimize the model's performance and ensure its effectiveness in meeting the commercial objectives of accurate risk tagging and classification |
| K5 How to design and deploy effective techniques of data analysis and research to meet the needs of the business and customers | Effective data analysis and research techniques will be applied to collect, preprocess, and analyze the news article dataset. This analysis will enable the development of the text classification model that meets the needs of the business and customers by accurately classifying news articles based on risk types. |
| K6 How data products can be delivered to engage the customer, organise information or solve a business problem using a range of methodologies, including iterative and incremental development and project management approaches | The project will employ an iterative and incremental development approach to design and deploy the text classification system. Project management methodologies will be utilized to ensure effective delivery of the data product, engaging stakeholders and solving the business problem of risk identification and classification. |
| K13 How to identify the compromises and trade-offs which must be made when translating theory into practice in the workplace | Throughout the project, compromises and trade-offs will be identified and addressed when translating theoretical concepts into practical implementation. Consideration will be given to balancing model complexity, computational resources, and the desired level of accuracy to meet the practical requirements of the workplace. |
| K14 The business value of a data product that can deliver the solution in line with business needs, quality standards and timescales | The project will focus on delivering a product, the text classification system, that provides a business solution in line with business needs, quality standards, and agreed-upon timescales. The value of the data product will be demonstrated through its ability to accurately tag and classify news articles based on risk types. |
| K23 The use of different performance and accuracy metrics for model validation in AI projects | Different performance and accuracy metrics will be utilized for model validation in the AI project. Metrics such as precision, recall, F1 score, and accuracy will be employed to assess the effectiveness and reliability of the text classification model in accurately classifying news articles. |
| K26 The scientific method and its application in research and business contexts, including experiment design and hypothesis testing | The project will follow the scientific method by applying research methodologies and experimental design to develop and evaluate the text classification model. Hypotheses will be formulated and tested to validate the effectiveness of the model in meeting the business objectives. |
| K28 How to communicate concepts and present in a manner appropriate to diverse audiences, adapting communication techniques accordingly | Concepts and findings will be communicated in a manner appropriate to diverse audiences. Presentations, reports, and documentation will be tailored to suit the technical and non-technical backgrounds of stakeholders, ensuring effective communication and understanding of the project's objectives and outcomes. |
| S2 Independently analyse test data, interpret results and evaluate the suitability of proposed solutions, considering current and future business requirements | Test data will be analyzed, and the results will be interpreted to evaluate the suitability and effectiveness of the proposed text classification solutions. Consideration will be given to current and future business requirements to ensure the solutions meet the identified needs. |
| S3 Critically evaluate arguments, assumptions, abstract concepts and data (that may be incomplete), to make recommendations and to enable a business solution or range of solutions to be achieved | Arguments, assumptions, and abstract concepts will be critically evaluated throughout the project to make informed recommendations and achieve the desired business solution. Data, even if incomplete, will be analyzed and assessed to provide robust conclusions and insights for effective risk tagging and classification. |
| S4 Communicate concepts and present in a manner appropriate to diverse audiences, adapting communication techniques accordingly | Concepts and findings will be communicated effectively to diverse audiences, adapting communication techniques accordingly. Presentations, reports, and documentation will be tailored to the specific needs and backgrounds of stakeholders, ensuring clarity and understanding of the project's progress and outcomes. |
| S5 Manage expectations and present user research insight, proposed solutions and/or test findings to clients and stakeholders. | User research insights, proposed solutions, and test findings will be presented to clients and stakeholders in a manner that manages expectations and fosters effective collaboration. The presentation of findings will facilitate decision-making processes and drive informed actions related to risk management based on the text classification system's outcomes. |
| S7 Work autonomously and interact effectively within wide, multidisciplinary teams | The project will involve working autonomously and collaborating effectively within a multidisciplinary team. Independent tasks will be managed individually, while teamwork will be crucial for areas such as data collection, model development, and system integration. Effective communication and collaboration within the team will be essential to ensure the project's success. |
| S9 Manipulate, analyse and visualise complex datasets | Complex datasets will be manipulated, analyzed, and visualized to extract meaningful insights and support the development of the text classification model. Advanced data analysis techniques will be applied to handle the complexity of the news article dataset and derive actionable information for accurate risk tagging. |
| S10 Select datasets and methodologies most appropriate to the business problem | The selection of datasets and methodologies for the project will be based on their relevance and suitability to address the business problem of risk classification. Careful consideration will be given to choose the most appropriate datasets and methods that align with the objectives and requirements of the text classification system. |
| S11 Apply aspects of advanced maths and statistics relevant to AI and data science that deliver business outcomes | Aspects of advanced mathematics and statistics relevant to AI and data science will be applied to ensure that the developed text classification model delivers desired business outcomes. Statistical techniques, algorithms, and mathematical models will be employed to optimize the model's performance and accuracy. |
| S15 Identify, develop, build and maintain the services and platforms that deliver AI and data science | The project will involve the identification, development, building, and maintenance of services and platforms necessary for delivering the AI-based text classification system. The implementation of these services and platforms will enable the effective deployment and utilization of the data product within the organization |
| S17 Consistently implement data curation and data quality controls | Data curation and data quality controls will be consistently implemented throughout the project to ensure the reliability and accuracy of the text classification system. Robust data cleaning, preprocessing, and validation techniques will be employed to address data quality issues and improve the overall performance of the system. |
| S18 Develop tools that visualise data systems and structures for monitoring and performance | Tools that visualize data systems and structures will be developed to monitor the performance and effectiveness of the text classification system. These tools will provide insights into the system's functioning, allowing for ongoing monitoring, optimization, and performance evaluation. |
| S22 Apply scientific methods in a systematic process through experimental design, exploratory data analysis and hypothesis testing to facilitate business decision making | The project will apply scientific methods in a systematic process, including experimental design and exploratory data analysis, to facilitate effective decision-making. The systematic approach will enable the identification and validation of risk tagging approaches and support the business in making data-driven decisions. |
| S24 Apply research methodology and project management techniques appropriate to the organisation and products | Research methodology and project management techniques appropriate to the organization and product will be applied throughout the project. This will ensure the project is well-structured, organized, and executed in a timely manner, meeting the defined objectives and delivering the desired outcomes. |
| S25 Select and use programming languages and tools, and follow appropriate software development practices | The project will involve the selection and utilszation of programming languages and tools relevant to AI and data science. The chosen programming languages and tools will be used to develop the text classification system and follow appropriate software development practices to ensure the quality and reliability of the system. |
| S27 Analyse information, frame questions and conduct discussions with subject matter experts and assess existing data to scope new AI and data science requirements | Information analysis, discussions with subject matter experts, and assessment of existing data will be conducted to identify and scope the AI and data science requirements for the project. These activities will help ensure that the text classification system meets the specific needs and challenges faced by the organization. |
| B2 Reliable, objective and capable of independent and team working | The project will demonstrate reliability, objectivity, and the ability to work both independently and within a team. The project team will collaborate effectively, and individual team members will demonstrate accountability, professionalism, and the capability to contribute independently to the project's success. |
| B6 Is comfortable and confident interacting with people from technical and non-technical backgrounds. Presents data and conclusions in a truthful and appropriate manner | Interacting with people from technical and non-technical backgrounds will be a key aspect of the project. Presenting data and conclusions in a truthful and appropriate manner will be essential to ensure effective communication and understanding among stakeholders with diverse backgrounds. Clear and concise explanations will be provided to facilitate comprehension and collaboration. |

## Conclusion

|  |  |
| --- | --- |
|  | Define the expected benefits achieved from using the project, the expected challenges faced during the life cycle of the project and the expected theoretical and empirical contribution. |

|  |
| --- |
| In conclusion, the proposed text classification project holds great potential for bringing several benefits to the organization. By leveraging AI and machine learning methodologies, the project aims to enhance the risk identification and classification process for news articles. The expected benefits include improved efficiency in identifying and categorizing risk-related articles, enhanced accuracy in risk tagging, and the ability to extract actionable insights from large volumes of news data.  However, the project is not without its challenges. Some anticipated difficulties include sourcing and curating a diverse and representative news article dataset, ensuring the reliability and quality of data sources, and addressing the inherent complexities of natural language processing and text classification. Additionally, the project team may encounter obstacles related to model performance optimization, feature engineering, and addressing the potential bias or interpretability concerns associated with AI algorithms.  From a theoretical perspective, the project is expected to contribute to the field of AI and data science by applying advanced statistical and mathematical methods to real-world commercial projects. The project's outcomes will further enrich the understanding of how AI and machine learning techniques can be effectively employed to meet business objectives, specifically in the context of risk identification and classification.  Empirically, the project aims to deliver a functional text classification system that meets the organization's needs and quality standards. The successful implementation of the project will provide tangible evidence of the value of AI-based solutions in the domain of supervision. Furthermore, the project's findings and insights can potentially contribute to further improvements into the tool by utilizing additional data and similar systems to supplement and compliment this tool.  Overall, by addressing the identified challenges and making meaningful theoretical and empirical contributions, the project aspires to deliver a robust text classification system that enhances risk management processes. The anticipated benefits, including improved efficiency, accuracy, and actionable insights, have the potential to drive informed decision-making, mitigate risks effectively, and contribute to the organization's overall success in managing associated risk identified within news articles. |

## Software Application Requirements

|  |  |
| --- | --- |
|  | Identify the software applications and tools to be utilised in the project, e.g. R Studio, SQL, Java, JavaScript, Python, Excel, PowerPoint, Word, Power BI, Tableau, QlikView, Qlik Sense, etc. |

|  |
| --- |
| Python, RStudio |

## Project Plan/Schedule

|  |  |
| --- | --- |
|  | Describe what the high-level timeline/schedule will be to plan, design, develop and deploy the project. Generally, by when do you expect this project to be finished? When should the key milestones be achieved? You can consider using a Gantt Chart. |

|  |
| --- |
| High-Level Project Timeline/ Schedule:  1. Project Planning Phase (1 week):  - Define project objectives, scope, and deliverables  - Identify project team members and stakeholders  - Develop a detailed project plan, including milestones, tasks, and dependencies  2. Design Phase (2 weeks):  - Gather and preprocess the news article dataset  - Conduct exploratory data analysis to gain insights into the data  - Design and implement a suitable text classification model architecture  - Perform feature engineering and selection  - Split the dataset into training, validation, and testing sets  3. Development Phase (4 weeks):  - Train and optimize the text classification model using appropriate machine learning techniques  - Evaluate the model's performance and iterate on the design as necessary  - Implement data pipelines and integration with the model  - Conduct rigorous testing and debugging  - Fine-tune the model to achieve desired accuracy and performance  4. Deployment Phase (2 weeks):  - Prepare the text classification system for deployment  - Integrate the system into the organization's infrastructure  - Conduct extensive testing and quality assurance  - Develop documentation and user guides  - Deploy the system to production environment  Key Milestones:  - Milestone 1: Project Planning and Kickoff - Expected completion: Week 1  - Milestone 2: Design Phase Completion - Expected completion: Week 3  - Milestone 3: Development Phase Completion - Expected completion: Week 7  - Milestone 4: Deployment and Finalization - Expected completion: Week 9  Please note that the provided timeline is a general estimate and can vary based on the complexity of the project, availability of resources, and unforeseen challenges that may arise during the project lifecycle. It is recommended to regularly monitor the project's progress, adjust the schedule as needed, and communicate any updates or delays to stakeholders to ensure a successful and timely completion of the project. |

## Apprentice Agreement

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Position/Role in the organisation | Signature  (e-signature or scanned wet signature accepted. Typed signature NOT accepted) | Date |
| Dylan Viswambaran | Data Scientist |  | 06/07/2023 |
| *By signing this proposal you are agreeing that you have read and will adhere to the*[*University's Ethical code of conduct*](https://policies.northeastern.edu/policy101/)*.* | | | |

## Mentor/Line Manager Approval

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Position/Role in the organisation | Signature  (e-signature or scanned wet signature accepted. Typed signature NOT accepted) | Date |
|  |  |  |  |

## Appendix

|  |  |
| --- | --- |
|  | Include the relevant supporting information in this section. Remove this section if not required |

­­