Mining Semantic Soft Factors For Agency-Client Responsiveness Prediction From Job Advertisement Descriptive Text.

Christopher Mathews | MSc Data Analytics | University of Huddersfield| 05.11.21.

Abstract
Big data analytics (BDA) offers little benefit to small and medium firms (SME) unless it can support the creation of profit (Chiang et al., 2018). A concurrence of missing empirical investigations to create economic value using BDA in SME's is acknowledged throughout information systems, information technology and organisational research. BDA being used to strengthen a firm's dynamic capabilities (DC) shows promise to be a key focal node for the realising economic value. Using the novel text mining method of Wang et al. (2020) this proposal seeks to discover if job advertisement descriptions can be mined for their semantic soft factors that might infer the probability of agency-client responsiveness. The method extracts the descriptive text element of job advertisements, from streams of real online data, to create cliques of related terms that may, or may not intuit agency responsiveness probability. Outcomes could inductively infer a BDA methodology with cross-domain utility with implications for SME mangers & executives. The proposed SME case-study would contribute to BDA literature, particularly categorised by its exploratory quantitative observations of DC development using semantic text mining.

1.0 Introduction
This proposal starts with a high-level overview of the type of research to be conducted, the aim, research questions (Kitchenham et al. (2009) poised and a detailed orientation of its originality within current literature. It continues with a discussion surrounding the proposed methodology and a breakdown of objectives to meet the outlined deliverables. The proposal closes with an analysis of risks and relevant issues.

1.1 Research Type
Quantitative, industry specific (recruitment) case study.

1,2 Research Aim
Evaluate if the outcomes of semantic text mining from job advertisements improves the average agency-client engagements.

1.3 Research Questions
1. *Can semantic soft factor mining from job description text create efficiencies in new business development tasks?*
2. *Can the Wang et al. (2020) soft factor mining model be applied cross-domain to have utility outside of credit risk?*

2.0 Literature Orientation
In 2021 the UK government released the National Artificial Intelligence Strategic Report revealing its continued intention to support firms of all sizes investing in their AI aptitude development (HM Government, 2021). This commitment should inspire AI projects up and down the country as progress towards an AI-enabled economy can move from early traction towards steady momentum. Underpinning the success of an AI-enabled economy will require knowledge sharing of how to create economic value from new technologies; duly of interest in the BDA literature (Günther et al., 2017; Chiang et al., 2018). If firms of all sizes are to succeed in the governments vision of an AI-enabled future, then small to medium firms (SME) must be able to overcome the principles of inertia (Mikalef et al., 2021) associated with developing a BDA strategy; a recent study by Kharade & Dong (2018) cautioned SMEs that over investing in IT projects which under-deliver comes with substantial financial risk. The question remains then, how do SMEs create a profit from BDA? An answer that has shown merit is using BDA outcomes at a managerial or executive level to make dynamic, data-driven decisions, empowering them to seize opportune, real time market advantages. Known as 'dynamic capabilities' (DC), the theory holds over time, introduced by Teece et al. (1997), and the call for further empirical testing in real settings is ubiquitous amongst BDA literature reviews concerning value creation (Davenport et al., 2012; Baker & Chasalow, 2015; Wamba et al., 2015; Günther et al., 2017). In the last 18 months 'the basket of 8' top IS journals have published quantitative BDA research that addresses the aforementioned gap from theory to application. Studies have shown competitive advantages in several business mandates, with a weighted quantity of papers focussing on marketing efficiencies (Dong & Yang, 2020; shin et al., 2020; molitor et al., 2020; Raguseo et al., 2021). Marketing efficiencies as a DC resulting from BDA activities is where this project proposes to add knowledge. Its orientation amongst BDA literature can be better visualised top down as follows:

- SME Big Data Analytics
    - Economic Value Creation
        - Organisation Dynamic Capabilities
            - Marketing Efficiencies

2.1 Domain Background

Recruitment, in the context of this proposal, is the HR practice of resourcing candidates as a billed service for clients seeking to fill open positions. Recruitment consultants can be valued for their deeply developed industry specific networks, which firms scarcely are exposed to through their in-house job advertisement efforts. Yet, networking of itself yields little value without clients to serve candidates to. New business development is a job prerequisite for recruitment consultants, a popular method for the consultant without distinguished clients is to browse job platforms and attempt to open a discovery conversation with the advertising company. Information pertaining to a firm's responsiveness for professional service consultation is limited by the fixed overview sections that advertising platforms mandate. Descriptive sections that offer more sentiment take time to access, accumulate and rank, a ready solution could be of important value to recruitment consultants looking to find efficiencies in their new business development labours. HR literature is voluminous in organisational research, however using the keywords 'recruitment' and 'big data analytics' offers no relevant results in the popular computer science directory IEEE Xplore. Along with the authors own domain interests, an exploratory case-study testing the interoperable replicability of the novel text mining method (Wang et al., 2020) triangulates an original gap in information systems and organisational literature.

3.0 Methodology Concept

This section starts by describing the cross-domain rationale and laymen overview of the experiment. Followed by a detailed breakdown of expected deliverables and outcomes from methodology design, development and deployment stages. These deliverables are next broken into individual baskets of milestones, objectives and tasks. Three select inspired the core structure of this section (Dawson, 2015; Wang et al., 2020; Zykov, 2021).

3.1 Analogical overview

Peer to peer lending and the recruitment sector appear to have little business overlap anecdotally. However, this study is not proposing that an individual's creditworthiness indicates their suitability for professional services marketing, rather though the same tools used to identify creditworthiness may be able to indicate responsiveness to relational development activities. It is this inductive inference that is proposed to explore experimentally. Can the semantic soft factor analysis model of Wang et al. (2020) have cross-domain utility in realising economic value? As seen in the early stated research questions.

3.2 Essential Problem Framing & Solution Understanding

In the Wang et al. (2020) study, the problem framed is simple: Identify good versus bad, applicants who pay on time versus applicants who default on financial commitments. The foremost basic requirement for a test of this class is a block of descriptive text that semantics can be derived from. In the original study loan applicants submitted additional comments alongside their applications on a peer-to-peer lending site. The problem was approached by developing a novel semantic text mining tool of which its algorithm could detect semantics after being previously trained by the researchers on a custom set of keywords related to creditworthiness. The outcomes were scored and aggregated by 'good versus bad' and benchmarked for statistical significance using historical loan granting or declining data. Disseminating the custom soft factor analysis model resulted in better credit risk prediction performed than other state-of-the-art semantic text mining tools used in the domain, Latent Dirichlet Allocation, in non-parametric tests ($p < 0.001$). These excellent results are an example of a BDA use-case in realising economic value through, in this case through increased prediction accuracy of loan granted.

4.0 Hypotheses

In Roman Zykovs' *Data Science: How to Monetize your Data* he cites a statistical heavyweight Fisher, "Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis." Fisher (1935, as cited in Zykov, 2021, p.208). The semantic soft analysis model is proposed by this project to be tested in an A/B setting is best understood by the following hypothesis:

H_0 Average responsiveness to agency-client engagement activities from recruitment consultants using a semantic soft factor ranking filter will be equal to those who do not.

H_1 Average responsiveness to agency-client engagement activities will be improved for recruitment consultants using a semantic soft factor ranking filter.

5.0 Design, Development & Deployment: Outcomes, Deliverables & Objectives

By way of recommendation from Dawson (2015), a breakdown of objectives and tasks are mapped to deliverables. Deliverable and outcomes can be found in this section. As the proposal works through the levels of work, specific tasks are described in an order delivered to minimise confusion. Visual elements have been used to illustrate tasks to objectives & objectives up to deliverable, supporting the overarching research aim and questions. One level below the proposed research aims, deliverables and outcomes of two distinct project phases, 'Design' and 'Development & Design' are offered first in Table 1.

Table 1. Research Aim Expected Deliverables & Outcomes

	Design	
	Deliverable	Outcome
1	Literature Review Report	*a.* Sound methodological & domain knowledge *b.* Knowledge acquisition gap awareness
2	Finalised Hypotheses	*a* . Testable metrics firmly defined

	Development & Deployment	
	Deliverable	Outcome
1	Working Soft Factor Text Mining Model	*a* . Trained keyword cliques *b.* Loading extracting texts *c.* Accurate semantic analysis
2	Report generated With Semantic Ranking Filter	*a.* Subjects able to view job advertisements ranked through the semantic responsiveness filter *b* . A/B test conducted
3	Results reported from A/B Test Pilot Case-Study	*a.* Statistical knowledge if model proves alternate hypothesis

5.1 Design Phase Outcomes & Objectives

Elaborating on Dawson (2015), Figure. 1 shows the work breakdown structure (WBS) for the 'Design' phase. Deliverable outcomes require a sound understanding of algorithm design. Considerations for alternate methods not covered in the Wang et al. (2009) study will be included for discovery. Training the model on a custom set of keywords requires background literature searching for the semantic of responsiveness. Figure 1. Shows focal nodes for a systematic review which this proposal deems essential as part of the risk mitigation process (See Appendix 2).

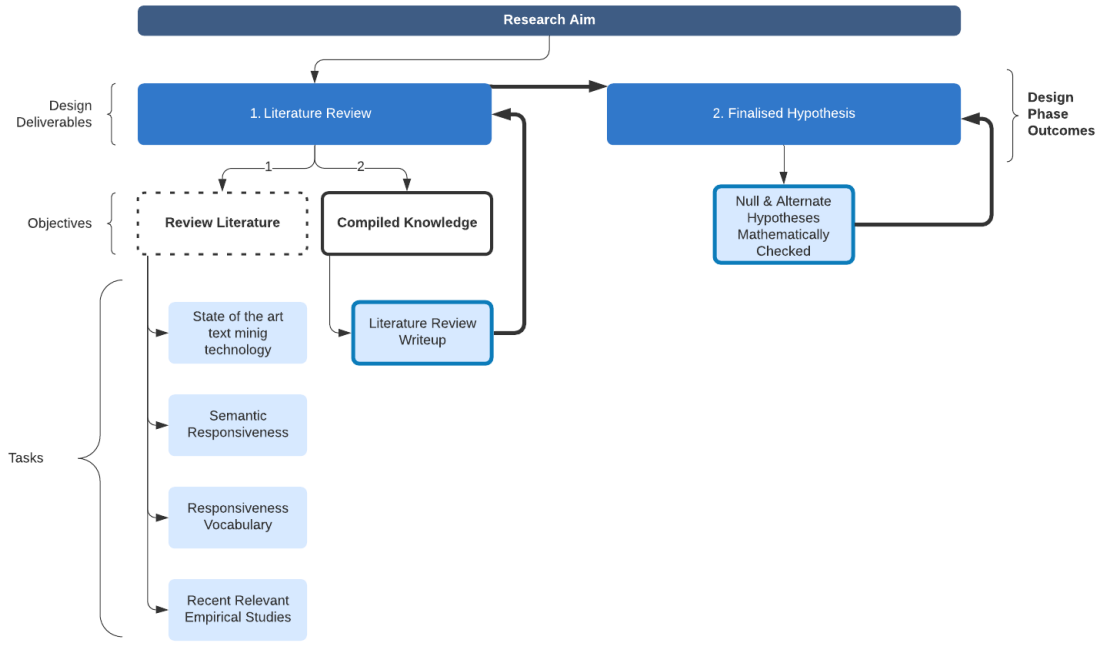


Figure 1. Design Phase WBS. Thicker arrows indicate critical path.

5.2 Development & Deployment Outcomes & Objectives

To comprehend the following section which describes the algorithmic structure, viewing Figure. 2 on the subsequent page is recommended. The outcomes of this phase include a working model that loads data, extracts meaning and generates a report, to then be used as the testable asset in an A/B test to investigate the hypotheses.

Figure 2. displays deliverable one in this phase will be the working model. This research is looking for cross domain utility of a proven working model, therefore it considers building the Wang et al. (2020) model in its truest representation essential for overall credibility and defining substance of the poised research questions. The following summarises key development phases using this case-study proposal for context.

The first stage of the model is to build a vocabulary of terms made up of single words and collocations. The proposed literature review aims to cover all bases with groups of words that might not be seemingly obvious from just scanning job descriptions solitarily. An embedded space is then created for terms to be mapped to and the semantic relationship captured. The reason for this is different advertisers will use different keywords that infer where they fall on a spectrum of responsiveness, but the semantic relationship between the different choice words will likely be similar. Resulting in groups of keywords, with a similar semantic 'clique'. Sequentially the model is trained on more keywords and collocations to refine & expand the cliques. Eventually resulting in a set of cliques, with semantic overview definitions which are known as the titled 'soft-factors'. A visualisation altered from Wang et al. (2020) for this case-study is provided in Figure 3 below.

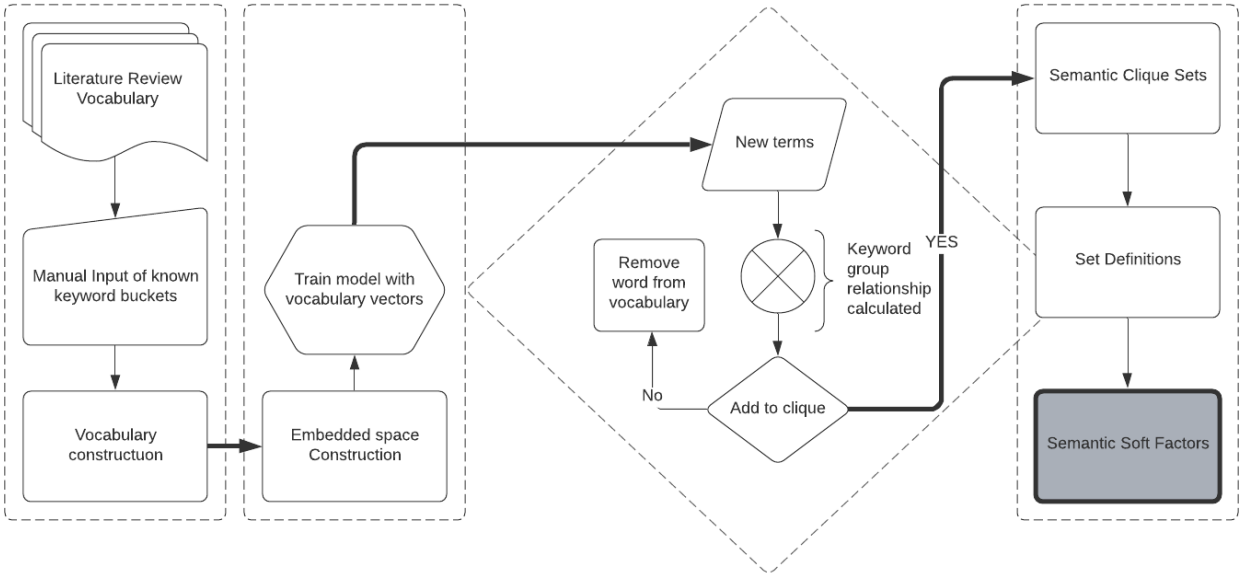


Figure 3. Development of Semantic Soft Factor Algorithm

After defining semantic soft factors, further data acquisition and preparation objectives are required to achieve the full set of outcomes from this deliverable. Considerations for API extraction and loading of data will be covered in the knowledge acquisition gap covered by phase one, the outcome at this stage will resemble a tuned model that loads website data from a job advertising platform and extract a semantic score of responsiveness, finally ranking said scores in descending order. Further coding, again identified in the first phase of deliverables, will be used to generate a readable report of the ranked job advertisements. Prior to conducting the A/B test, as the testable metric is mostly complete, professional domain consultation will be sought to examine initial results before continuing. The third deliverable in this phase has two spokes, conducting the experiment and the interpretive discussion and conclusions. Tasks for the former involve setting up the environment for the samples, illustrated in Figure. 4 (top of page after next), collecting results and processing of raw statistical values.

5.2Development & Deployment Outcomes & Objectives

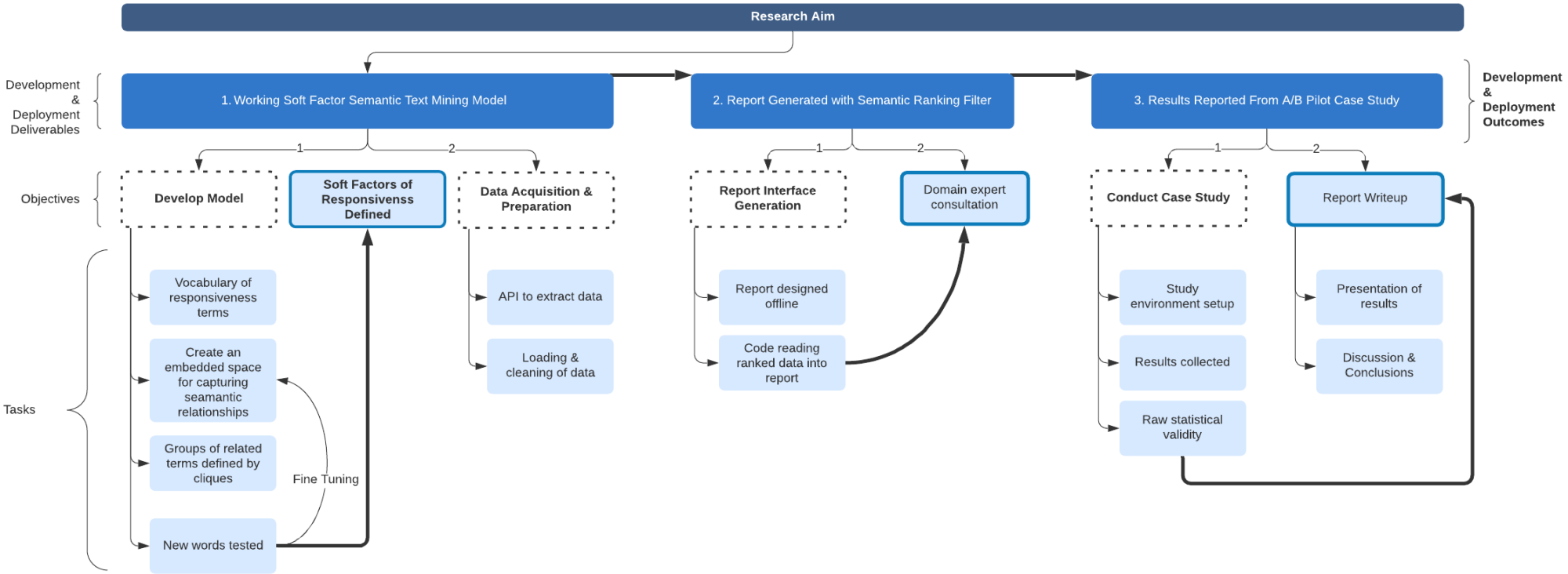


Figure 2. Development & Deployment WBS. Thick arrows indicate critical path

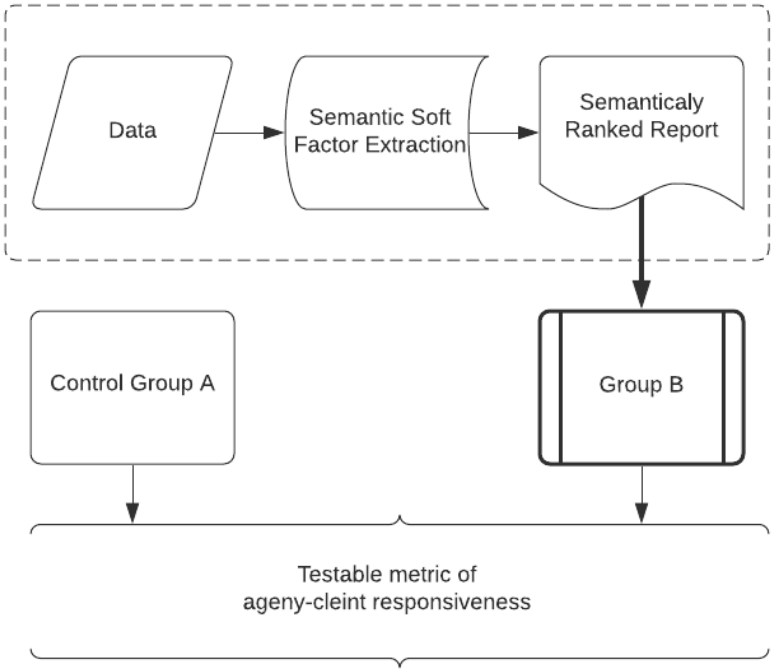


Figure 4. A/B Test Environment Setup

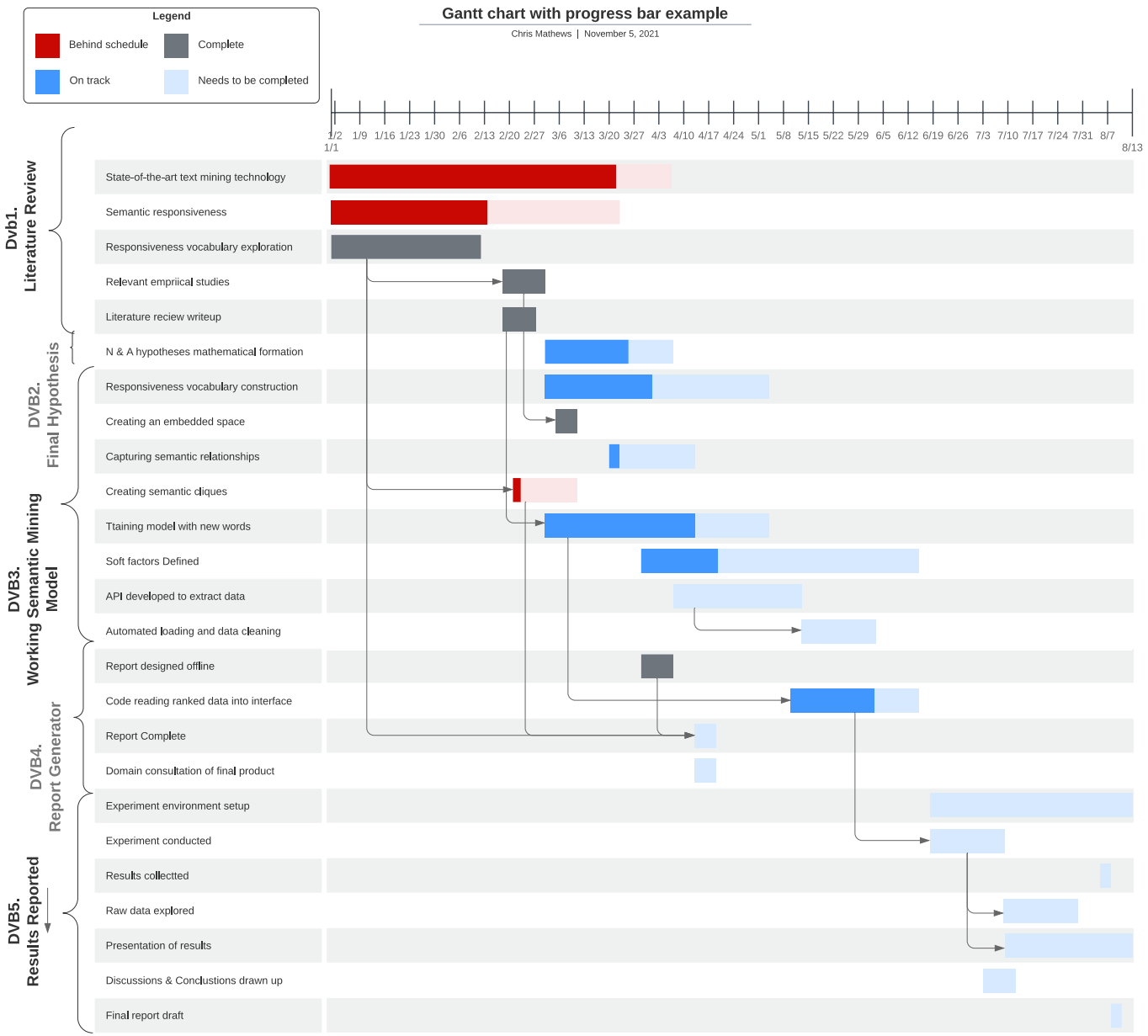
6.0 Project Timeline Management

Table 2. shows deliverable timelines broken into phase objectives. Individual Task breakdowns are visualised in the appending Gaant chart (Appendix 1). The chart is adapted with Computer Science project planning (Dawson, 2015) in mind. Indicating areas where slack may apply but is limited in showing constraints. A notable limitation of this project plan is the delimiting of a critical path diagram. However, it is arguably not required due to the sequential design of the deliverables; one cannot begin until the other is finished. An attempt to display project constraints has been visualised on the Gaant chart which would likely suffice for a master's level project. For working demonstration purposes, objects have active plotting shown in the chart.

Table 2. Deliverable Objectives with Expected Duration in Weeks.

	Deliverable	Obectives	Expected Duration
1	**Literature Review Report**	Review Literature	*6 weeks*
		Compile Knowledge	*2 weeks*
2	**Finalised Hypotheses**	Finalised Hypothesis Checked for Mathematical Soundness	*1 week*
1	**Working Soft Factor Text Mining Model**	Develop Model	*9 weeks*
		Data Acquisition & Preparation	*4 weeks*
2	**Report generated With Semantic Ranking Filter**	Report Interface Generation	*1 week*
		Domain expert Consultation	*< 1 week*
3	**Results reported from A/B Test Pilot Case-Study**	Conduct Case Study	*5 weeks*
		Report Writeup	*2 weeks*
		Total	**>31 weeks**

APPENDIX 1 – GAANT CHART



APPENDIX 2 – EXAMPLE RISK ASSESSMENT DEMONSTRATING A SELECTION OF POTENTIAL RISKS

ID	Category	Description	Consequence	Probability (1-3)	Impact (1-3)	PI Score	Risk mitigation plan	Risk owner
1	Scope	Literature review & knowledge acquisition objectives bring forth unseen tasks	Scope of project too large or complex, workarounds will stray from research aims.	3	3	9: High	Reseasrch Aims will be air tight to guide the literature review. Constrain knowledge to neccessities.	Christopher Mathews
2	Technical	Model performs worse than expected	Extra time spent addressing performance issues	2	3	6: Medium/High	Consult model design plan with a field expert before development begins.	Christopher Mathews
3	Project	Tasks take longer than plotted.	Objectives fall behing and deliverables delayed. Knock on effect felt through whole timeline.	3	3	9: High	Real time Gaant chart will be used to plot tasks early. Project manager will be consulted for opening a buffer window.	Christopher Mathews
4	PLES	Advertising platform does not allow extraction of the data required.	An agreement for research might need to be sought, or a new platform option explored.	2	2	4: Medium/Low	A backup platform will be selected, permission if not explicitly granted, will be requested in advance.	Christopher Mathews
5	PLES / Technical	Accessibility issues with end report for use in A/B test	Users can't interpret report and therefore cannot use it. Report does not open and cannot be used.	1	3	3: Low/Medium	Report design will have accessibiity considerations before development and tested before experiment.	Christopher Mathews
6	Business	Model gives incorrect advice	The test group do not perform any better than the control as the information recieved is incorrect.	2	3	6: Medium/High	A pilot test will be run in advance to make sure model is working correcntly.	Christopher Mathews

Bibliography

1. Baker, E., & Chasalow, L. (2015). Factors contributing to business intelligence success: The impact of dynamic capabilities [Paper presentation]. *Twenty-First Americas Conference on Information Systems, Puerto Rico,* https://www.researchgate.net/profile/Lewis-Chasalow/publication/284156909_Factors_Contributing_to_Business_Intelligence_Success_The_Impact_of_Dynamic_Capabilities/links/564c8eff08ae020ae9faba79/Factors-Contributing-to-Business-Intelligence-Success-The-Impact-of-Dynamic-Capabilities.pdf

2. Chiang, R. H. L., Grover, V., Liang, T., & Zhang, D. (2018). Special Issue: Strategic Value of Big Data and Business Analytics. *Journal of Management Information Systems, 35*(2), 383-387. 10.1080/07421222.2018.1451950

3. Davenport, T. H., Barth, P., & Bean, R. (2012). How 'big data' is different. *MIT Sloan Management Review, 54*(1), 43.

4. Dawson, C. W. (2015). *Projects in computing and information systems: a student's guide* (Third ed.). Pearson Education.

5. Edwards, A. W. F. (2005). R.A. Fischer, statistical methods for research workers, first edition (1925). (First ed., pp. 856-870)10.1016/B978-044450871-3/50148-0

6. Fosso Wamba, S., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics, 165*, 234-246. 10.1016/j.ijpe.2014.12.031

7. Gentsch, P. (2018). *AI in Marketing, Sales and Service: How Marketers without a Data Science Degree can use AI, Big Data and Bots* (1st 2019. ed.). Springer International Publishing.

8. Günther, W. A., Rezazade Mehrizi, M. H., Huysman, M., & Feldberg, F. (2017). Debating big data: A literature review on realizing value from big data. *The Journal of Strategic Information Systems, 26*(3), 191-209. 10.1016/j.jsis.2017.07.003

9. HM Governement. (2021). *National AI Strategy (Report).* (No. 1.2). United Kingdom: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1020402/National_AI_Strategy_-_PDF_version.pdf

10. Karhade, P. P., & Dong, J. Q. (2020). Information Technology Investment and Commercialized Innovation Performance: Dynamic Adjustment Costs and Curvilinear Impacts. *MIS Quarterly,*

11. Kitchenham, B., Pearl Brereton, O., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering – A systematic literature review. *Information and Software Technology, 51*(1), 7-15. https://doi-org.libaccess.hud.ac.uk/10.1016/j.infsof.2008.09.009

12. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.

13. Mikalef, P., van de Wetering, R., & Krogstie, J. (2021). Building dynamic capabilities by leveraging big data analytics: The role of organizational inertia. *Information & Management, 58*(6)10.1016/j.im.2020.103412

14. Molitor, D., Spann, M., Ghose, A., & Reichhart, P. (2020). Effectiveness of Location-Based Advertising and the Impact of Interface Design. *Journal of Management Information Systems, 37*(2), 431-456. 10.1080/07421222.2020.1759922

15. Raguseo, E., Pigni, F., & Vitari, C. (2021). Streams of digital data and competitive advantage: The mediation effects of process efficiency and product effectiveness. *Information & Management, 58*(4), 103451. 10.1016/j.im.2021.103451

16. Teece, D. J., Pisano, G., & Shuen, A. (1997). Dynamic capabilities and strategic management. *Strategic Management Journal, 18*(7), 509-533. 10.1002/(SICI)1097-0266(199708)18:7<509::AID-SMJ882>3.0.CO;2-Z

17. Wang, Z., Jiang, C., Zhao, H., & Ding, Y. (2020). Mining Semantic Soft Factors for Credit Risk Evaluation in Peer-to-Peer Lending. *Journal of Management Information Systems, 37*(1), 282-308. 10.1080/07421222.2019.1705513