INITIAL RESEARCH PROPOSAL FORM

(also referred to as 'Statement of Intent Form')

To be submitted by the researcher to the Institute Research Sub-Committee (IRC)

Research Title:

Using NLP derived Telemetry and Predictive Analytics to Refine Business Decisions in Logistics Operations.

Institute name

Institute of Information and Communication Technology,

Course / Programme:

Bachelor of Science (Honours) in Business Analytics

Level and year of study

MFQFL 6, Second Year

Main area of study being proposed:

This study explores the benefits and advances that Predictive Analytics bring to Business Decision Making. Such techniques require a vast data set to power a machine learning algorithm that could be applied to current parameters and in turn provide insight to the decision to be made.

The aim of the study is to take the algorithm created and apply it to several situations were businesses are required to make decisions whilst taking several logistics-based factors. What are the possible delays between each way point? Would the deliverable arrive on time? If the deliverable will not make it on time, should the business purchase another one from a supplier that is closer to ensure on-time delivery or should they incur the fines?

Commented [AG1]: considerable

Commented [AG2]: trained model

All of these situations could arise during daily operations and require insight to make decisions that would push the business forward. The study aims to bring greater insight to the decision-making process should such situations arise. Once the prototype is complete, the system would be able to ingest real time data and produce possible outcomes and what their effects would do to the Business in terms of possibility of profit, possible fines incurred and whether on time delivery would be possible.

The machine logic algorithm requires a dataset. This dataset would be constructed by applying natural language processing to various forms of media that were either procured through RSS feeds or web scraping. Powered by Python, the NLP would look for specific tags to extract key components. The study will be focusing on Air Freight since this method of transportation is used for cases where the deliverables in question are needed as soon as possible at the destination.

Name of Researcher:	**Researcher's I.D. Number:**
Matthew De Giorgio	0047796M
Signature of Researcher	**Date of submission of Form**
	5th April 2019

Name of Tutor (or Recommended Tutor):

Mr. Alan Gatt

Personal Motivation for the Choice of Research Theme.

Coming from a family background of pharmaceutical importation and then starting to work in a leading technology consulting company, I felt that this research area was a perfect fit. Amidst daily operations, our director of logistics would encounter many situations were quick decisions had to be made. If a product shipment got delayed, we would have to decide whether we leave it alone and incur the late delivery fine or we purchase another shipment from a different supplier that would guarantee the product would arrive on time at a much higher cost. By the time the Net Profit calculations for each situation were completed, the time window for a positive outcome would have closed.

Having experienced this situation from a personal level, I was intrigued to explore the possible benefits that this technology could bring if it was to be implemented in the local market space. Seeing that large foreign companies have similar systems in place to make such situations as easy as possible, why not try to make a similar system and apply to a small company?

With this context, the purpose of this research is to create a prototype and assess whether it's feasible to be implemented into the local companies to reduce the strain on resources and management whilst increasing profitability in the highly competitive local market and perhaps bring some new insight and value to the family's business.

Outline of Key Literature and Theoretical Framework or Propositions.

Natural Language Processing

Current day NLP has many technologies that could be applied to many current problems. The main technology utilised in this study is Text Processing by use of the NLTK python library to aid Information extraction. This technology is used to source meaningful information from large amounts of unstructured text (D. Eggers et al,2017). This could be done through Part of Speech Tagging and token system. Important terms ('Tags') are set as parameters in the script. The program will go through the information dump and Count / Highlight all occurrences of the tags. Researchers applied these techniques to filter through police reports and related news article to identify key information such as weapons, vehicles, locations and people with high precision (C.D.Manning,2011).

Web Scraping

Web scraping is used for many different purposes, from data gathering to web page indexing. As mentioned beforehand, the first very first web scraper was, in fact, a web crawler based on the Perl programming language. The sole purpose of the World Wide Web Wanderer (WWWW) was to measure the size of the World Wide Web and to generate an index (called Wandex) in 1993 *(M. Grey,1996)*. The primary function of the Web crawler was to build the indexes for search engines since in the early days of the Web there were not that many websites and required website administrators to collect the links and enter them manually into a search engine Index.

Beautiful Soup is one of the most common approaches to web scraping since it does not depend on an API interface (Application programming interface) but instead, it parses web page content directly from the HTML container. Designed for Python, the library receives continuous updates and new functionalities every year

Significance of the Study.

The main focus of the study and the prototype is to predict the result of a shipment with a certain degree of accuracy using past and live data. A dashboard will be created to show the key stats of a shipment, the Net profit if it were to arrive on time and the net profit if it were to arrive late. Other Suppliers that also sell the product will be listed including an estimated purchase price, the last possible date to order for the shipment to arrive on time and the Net Profit if that supplier was to be used.

Another focus is to derive the main factors that are involved during transportation which could be used to predict the outcome of the shipment (the expected arrival vs the actual arrival time).

All data sets used will be based on European Air Traffic since that the operating zone of many local companies. If the data sets were to be expanded to include Air Traffic within the United States, future research could be conducted to answer the following questions:

I) Would the algorithm work with Transocean air-freight?
II) Could the algorithm evolve to predict the most efficient purchasing strategy if a big enough database of suppliers was linked to it?

The prototype may bring value to any business that exports and imports to sustain business deals and tenders that require goods to arrive on particular days.

Hypotheses and/or Research Question/s

I) Can Shipments be predicted accurately using past data?

II) Would further broadening of the data sources increase the accuracy of the algorithm?

Commented [AG3]: I would limit the research questions to one or two. Since if you ask them, you have to answer them :)

Target Participants and Research Methods for Data Collection and Analysis

1) Natural language processing would be used on RSS feeds and on web scraping data (using Beautiful Soup for Python) to create a dataset of events and the time delay they put on air freight.
2) A Microsoft SQL database would be built to house the data. Suppliers, products, air freight providers will also be included in the database.
3) Analysis would be conducted on the dataset to find patterns, and key statistics (fastest way from point A to Point B, cost per kg for transport)
4) An aspect of probability would have to be included to factor in the chances of delays during particular times of the year, particular locations (Ex: Snow in the northern Regions of Europe)
5) The results be visualized using Tableau or PowerBi to compare the generated models and graphs.
6) Measuring of the results and statistical tests to validate the algorithm and to show why such results are being given.
7) The results would be compared to actual destination arrival times.

Ethical Considerations.

Refer to *guidance points below. You are also additionally required to read MCAST Document 074 'Research Ethics Policy and Procedure' that is available on the College website via link http://www.mcast.edu.mt/MainMenu/Full-TimeCourses/Rules,PoliciesandRegulations.aspx*

1. *Research shall be conducted in such a manner so as to avoid any psychological and physical harm to humans and animals and financial damage to organizations*
2. *Only the supervisor and examiners will have access to any data gathered.*
3. *Participants will remain free to withdraw from the study at any time without having to provide any reason. In the case of withdrawal, all the records and information collection will be deleted.*
4. *The participant, who is the sole proprietor of the data provided, is granting that such data would be processed for this study purposes only.*
5. *The data collection process will be a transparent process.*
6. *All transcriptions and/or electronic recordings reflecting the data collected, once exhausted, are to be deleted*
7. *Confidentiality, anonymity and data protection procedures are to be ethically abided by.*
8. *The researcher would provide a soft copy of the study to the participant, if required.*

Enter details here regarding possibility of issues regarding confidential personal data:

Note: Participants refers to all Primary, Secondary and Key Stakeholders
Information collected in the proposed research study will be considered as information that the participants have "disclosed in a relationship of trust" and so no information will be divulged without their permission. Anonymity means that the research cannot link individual responses with the participants' identities. All the information collected during the study will be kept in a multi-factor authentication secured cloud storage facility. All information
processed during the study serves only for the purpose of the study itself and therefore once the research is complete, all transcripts and any information reflecting the data collected are to be destroyed. All the information collected during the study will be kept in a biometrically secured cloud storage facility.
Attached to this SOI is a model Consent form that will be issued before any research activities are to be initiated.

Enter details here regarding possibility of physical harm:

The methodology pertained by this study does require the use of animals for the scope of the study, neither will there be any physical harm to any interview attendees or respondents of the questionnaire. Furthermore, no Personal Protective Equipment (PPE) will be needed as the questionnaires and interviews will be dispensed and carried out in a safe office environment to eliminate the possibility of physical harm.

Enter details here regarding possibility of moral harm:

Maintenance of the basic ethical principles is to be upheld to highest of standards during the data collection process
including doing good, protecting the autonomy, wellbeing, safety and dignity of all participants. The researcher will maintain objectivity and will persist to avoid any possible psychological, spiritual or cultural misunderstandings with
participants. Participants always have the right to refuse to answer any questions even though the questionnaire and interview are designed to not involve any professional and emotional risks.

Enter details here regarding the possibility of business harm:

All research findings will be processed for the purpose of the research only and nothing more and will remain confidential and therefore, participants should not encounter in any way any competitive disadvantage as an the outcome of the research, nor will there be any form of harm to any businesses locally and abroad.
All Data will be held and view only by the Researcher and no one else. As previously described, Secure online storage
will be heavily utilised to reduce the chance of any data loss or breach.
Anonymity and data protection procedures are always to be upheld and so the proposed research methodology will
not divulge or reveal any confidential trade secret or protected data.

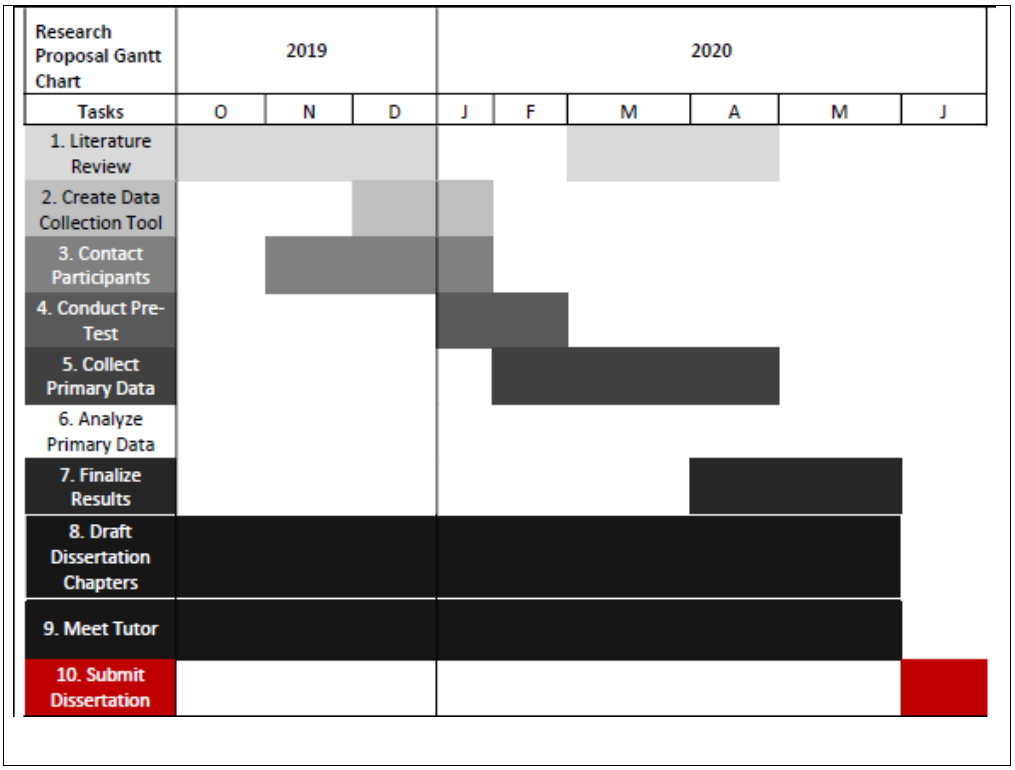
Anticipated Contributions of the Study.

The study aims to create a system that could contribute to a business in day – to – day operations. Locally quite a few businesses still use old and outdated systems that could the management critical time that could be used in important business decisions.

The system utilizes new technologies that have not been tested locally but the possibility for advancement is present.

Dissertation Project Plan.

The research project will span over nine months with the following chart depicting how research tasks will be split up accordingly. The Timeline is tentative for not all possible setbacks can be taken into consideration at the time of writing.



List of Key References:

Bringsjord, Selmer, Govindarajulu and Sundar, N. (2018). *Artificial Intelligence (Stanford Encyclopedia of Philosophy)*. [online] Plato.stanford.edu. Available at: https://plato.stanford.edu/entries/artificial-intelligence/#HistAI [Accessed 1 Mar. 2019].

D. Eggers, W., Malik, N. and Graciee, M. (2019). *Using AI to unleash the power of unstructured government data*. [online] Deloitte Insights. Available at: https://www2.deloitte.com/insights/us/en/focus/cognitive-technologies/natural-language-processing-examples-in-government-data.html [Accessed 24 May 2019].

D.Manning, C. (2011). *Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?*. [ebook] Stanford: Department of Linguistics , Stanford University. Available at: https://nlp.stanford.edu/pubs/CICLing2011-manning-tagging.pdf [Accessed 20 Apr. 2019].

Gray, M. (1996). *Internet Growth and Statistics: Credits and Background*. [online] Mit.edu. Available at: http://www.mit.edu/~mkgray/net/background.html [Accessed 3 Jun. 2019].

Hamaz, K. and Benchikha, F. (2017). A novel method for providing relational databases with rich semantics and natural language processing. *Journal of Enterprise Information Management*, [online] 30(3), pp.503-525. Available at: https://emeraldinsight.com/doi/full/10.1108/JEIM-01-2015-0005 [Accessed 2 Jun. 2019].

Iriberri, A. and Leroy, G. (2007). Natural Language Processing and e-Government: Extracting Reusable Crime Report Information. *2007 IEEE International Conference on Information Reuse and Integration*. [online] Available at: https://ieeexplore.ieee.org/document/4296624 [Accessed 11 May 2019].

Loira, S. (2018). *API Reference — TextBlob 0.15.2 documentation*. [online] Textblob.readthedocs.io. Available at: https://textblob.readthedocs.io/en/dev/api_reference.html#textblob.blob.TextBlob.sentiment [Accessed 2 Jun. 2019].

M. Turing, A. (1950). *OMPUTING MACHINERY AND INTELLIGENCE, Mind,*. 59th ed. Oxford: Oxford University Press, pp.433-460.

Manning, C. and Schütze, H. (1999). *Foundations of Statistical Natural Language Proccesing*. Cambridge (Massachusetts): MIT Press.

Murphy, K. (2013). *Machine learning*. Cambridge, Mass.: MIT Press.

Patel, H. (2018). *How Web Scraping is Transforming the World with its Applications*. [online] Towards Data Science. Available at: https://towardsdatascience.com/https-medium-com-hiren787-patel-web-scraping-applications-a6f370d316f4 [Accessed 1 Jun. 2019].

Pennington, J., Socher, R. and D.Manning, C. (2014). *GloVe: Global Vectors for Word Representation*. [online] Aclweb.org. Available at: https://www.aclweb.org/anthology/D14-1162 [Accessed 4 Jun. 2019].

Raina, R., Madhavan, A. and Y. Ng, A. (2009). *Large-scale Deep Unsupervised Learning using Graphics Processors*. [ebook] Stanford: Stanford University ,Standford. Available at: http://robotics.stanford.edu/~ang/papers/icml09-LargeScaleUnsupervisedDeepLearningGPU.pdf [Accessed 3 Jun. 2019].

Russell, S. and Norvig, P. (2009). *Artificial intelligence*. Upper Saddle River, N.J.: Prentice Hall.

T. Mueller., E. (2006). *Commonsense Reasoning*. 2nd ed. Morgan Kaufmann Publishers.

This section is to be filled in by the representative of the Institute Research Sub-Committee prior to forwarding of this Form to the 'MCAST Research Ethics Committee' for final ethics approval:

Nature of ethical consideration	***Outcome (*)***	***Comments***
Consideration of possibility of issues regarding confidential personal data:		
Consideration of possibility of physical harm		
Consideration of possibility of moral harm		
Consideration of possibility of business harm		

(*) Legend to record outcome by Institute Research Sub Committee:

A *– Ethical considerations have been* ***addressed appropriately*** *by Researcher;*
B *– No (****Nil****) relevant ethical considerations are applicable under purpose of study as described by Researcher.*
C *– Ethical consideration have* ***not been addressed appropriately*** *by Researcher;*
D *– Applicable ethical consideration have* ***not been considered*** *by Researcher.*

Details of Representative to the 'Institute Research Sub-Committee.	
Name	Signature
Designation	Date